

Федеральное государственное автономное образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
Базовая кафедра интеллектуальных систем управления

УТВЕРЖДАЮ

Заведующий кафедрой

_____ Ю.Ю. Якунин

«11» июня 2018 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 Системный анализ и управление

Идентификация многомерных дискретно-непрерывных процессов по выборке
наблюдений с выбросами

Руководитель

подпись, дата

должность, ученая степень

А.А. Корнеева

инициалы, фамилия

Выпускник

подпись, дата

М.А. Денисов

инициалы, фамилия

Красноярск 2018

РЕФЕРАТ

Темой выпускной квалификационной работы является «Идентификация многомерных дискретно-непрерывных процессов по выборке наблюдений с выбросами». Количество страниц – 62, иллюстраций, используемых в тексте – 17, таблиц с информацией – 2. В работе описано 29 формул, а также использовано 50 источников литературы.

Ключевые слова: идентификация, непараметрическая регрессия, робастный анализ, цензурирование данных, выбросы.

Рассматриваются вопросы непараметрической идентификации процессов, что является актуальным на современном этапе развития технологических систем в производстве и многих других отраслях жизнедеятельности человека. Выборка наблюдений содержит в себе выбросы, что уменьшает точность аппроксимации объекта. Предложены алгоритмы, позволяющие нивелировать влияние выбросов. Принцип и математическое описание алгоритмов неодинаково, что позволяет провести их сравнительный анализ.

Полученные результаты показывают эффективность работы предложенных алгоритмов, что выражено в значениях относительных ошибок аппроксимации. Выявлены некоторые особенности алгоритмов, в частности, для задачи ремонта данных с использованием робастного алгоритма при восстановлении значений объекта лучше использовать не частный случай формулы Надарая-Ватсона, а ее робастный аналог.

Исследованные в работе алгоритмы нашли свое применение при моделировании системы оценки стоимости недвижимости с применением реальной выборки наблюдений однокомнатных квартир города Красноярска.

СОДЕРЖАНИЕ

Введение.....	5
1 Идентификация многомерных дискретно-непрерывных процессов	7
1.1 Задача моделирования	7
1.2 Машинное обучение	11
1.3 Классификация. Прогнозирование	13
1.4 Параметрическое моделирование	15
1.5 Непараметрическое моделирование	17
1.6 Анализ данных. Выбросы. Робастный анализ.....	18
1.7 Выводы по главе 1	22
2 Алгоритмы обработки данных с выбросами.....	24
2.1 Классификация методов работы с выбросами	24
2.2 Робастная оценка регрессии №1	25
2.3 Робастная оценка регрессии №2	27
2.4 Непараметрический алгоритм цензурирования данных с выбросами...30	
2.5 Ремонт данных выборки наблюдений с выбросами	31
2.6 Выводы по главе 2	32
3 Вычислительные эксперименты	33
3.1 Вычислительный эксперимент с использованием робастного алгоритма №2.....	33
3.2 Вычислительный эксперимент с использованием алгоритма цензурирования данных	38
3.3 Вычислительный эксперимент по ремонту данных	41
3.4 Сравнение результатов работы алгоритмов	43
3.5 Выводы по главе 3	45
4 Задача оценки стоимости недвижимости.....	47
4.1 Постановка задачи	49
4.2 Непараметрическая модель	50
4.3 Вычислительный эксперимент.....	51

4.4 Информационная система	54
4.5 Вывода по главе 4	56
Заключение	57
Список использованных источников	59

ВВЕДЕНИЕ

Задача моделирования систем как никогда востребована в современном мире. Для того, чтобы развивать и управлять технологической, медицинской, экономической и другими отраслями человеческого производства требуются специальные методы. Выпускная квалификационная работа направлена на решение задачи моделирования, что подчеркивает ее актуальность.

В рамках данной работы будут применяться методы идентификации систем. В зависимости от количества априорной информации идентификация подразделяется на два типа:

- в «широком» смысле, когда априорной информации об объекте исследования недостаточно, чтобы построить модель с точностью до вектора параметров, в таких случаях пользуются непараметрическими методами моделирования;

- в «узком» смысле, когда параметрическая структура объекта исследования известна с точностью до вектора параметров и априорных данных достаточно, чтобы воспользоваться параметрическими методами идентификации.

При выполнении работы предполагается исследовать объект, априорной информации о котором недостаточно. В связи с этим, непараметрический подход к построению модели наиболее целесообразен.

Очевидно, что выборка наблюдений является неотъемлемой частью в процессе конструирования модели. В реальной жизни довольно часто приходится иметь дело с дефектными данными, которые, к примеру, содержат в себе выбросы. Поэтому задача совершенствования качества данных несет в себе немаловажное значение и является не менее актуальной, чем построение модели процесса.

Цель работы заключается в повышении точности идентификации процессов дискретно-непрерывного типа по выборке наблюдений с выбросами.

Задачи, возникающие в ходе реализации поставленной цели, следующие:

– поиск и исследование наиболее значимой литературы в выбранной области исследований;

– реализация и исследование непараметрического алгоритма идентификации для объектов одномерного и многомерного типа на основе оценки Надарая-Ватсона;

– реализация и исследование робастного алгоритма идентификации, а также алгоритма цензурирования выборки наблюдений с выбросами. Решено было взять алгоритмы, различные по своему принципу работы. Основоположником одного из этих подходов является ученый и профессор по статистике Питер Дж. Хьюбер;

– реализация алгоритма восстановления очищенных от выбросов данных;

– непараметрическая идентификация объекта дискретно-непрерывного типа, характеризуемого реальными данными. В процессе идентификации будут применены алгоритмы «очистки» данных от выбросов, упомянутые выше. При идентификации используется выборка наблюдений, характеризующая параметрами однокомнатных квартир города Красноярска.

В качестве инструментов, реализующих поставленные задачи, служат методы математического моделирования, в частности, имитационные методы (с использованием электронных вычислительных средств), а также методы математической статистики, что позволяет глубже изучить объект исследования.

1 Идентификация многомерных дискретно-непрерывных процессов

1.1 Задача моделирования

В реальной жизни довольно часто возникает потребность в исследовании какого-либо явления или процесса с целью детального изучения его поведения и свойств. Такая необходимость актуальна для промышленной, научной и многих других сфер деятельности. Однако, работать напрямую с исследуемыми процессами или явлениями не всегда возможно. Это обуславливается различными факторами, где в качестве основных можно выделить:

- высокую стоимость проводимых экспериментов над реальным объектом;
- необходимость иметь большое количество времени для осуществления эксперимента;
- проведение эксперимента в реальных условиях может быть опасно для жизни;
- проведение эксперимента в реальных условиях невозможно из-за уникальности исследуемой системы (система существует в одном экземпляре, либо недоступна).

В качестве примеров можно отметить проектирование космических аппаратов, жидкостных ракетных двигателей [1, 2], где абсолютно нецелесообразно проводить эксперименты над реальными объектами из денежных соображений. Или же исследование новых, мало изученных явлений природы. К примеру, открытие гравитационных волн в 2016 году [3] осуществилось благодаря наблюдениям над слиянием двух черных дыр. По результатам наблюдения была построена модель, с помощью которой в дальнейшем будут исследовать поведение и свойства данного явления. Очевидно, исследовать данное явление в реальных условиях не представляется возможным, в связи с этим задача построения его математической, физической модели является перспективной.

Следовательно, в подобных ситуациях целесообразно пользоваться моделями реальных объектов или явлений. Советов Б. Я. и Яковлев С. А. в своей книге [4] определяют моделирование «как замещение одного объекта другим с целью получения информации о важнейших свойствах объекта оригинала с помощью объекта-модели».

Для того, чтобы построить модель, пользуются различными методами. В общем случае, методы моделирования систем можно условно разделить на физические и математические (Рисунок 1) [5].



Рисунок 1 – Классификация видов моделирования

Физическое моделирование предполагает использование самой исследуемой системы или системы, с похожей физической природой. В качестве примера можно упомянуть исследование свойств летательных аппаратов в аэродинамической трубе.

Под математическим моделированием понимают процесс установления соответствия полученной математической модели некоторой реальной системе, с целью изучения этой модели для того, чтобы получить характеристики исследуемой системы. Используя аналитический способ, для построения модели, исследуемый процесс записывается в виде математических выражений (алгебраических, интегро-дифференциальных и т.п.), которые в дальнейшем могут быть исследованы с использованием следующих методов:

- аналитический метод (поиск в общем виде явных зависимостей искомых характеристик);
- численный метод;
- качественный (без использования решения в явном виде можно найти свойства решения).

Математическое моделирование с использованием компьютера как никогда актуально и широко распространено в настоящее время. В зависимости от используемого при построении модели математического аппарата, а также способа организации вычислительных экспериментов можно строить модель следующими способами:

- численно. То есть с использованием методов вычислительной математики, где необходимо численно решить математические уравнения при заданных начальных условиях и значениях определенных параметров;
- с использованием имитационного моделирования. То есть, воспроизведение на ЭВМ процесса функционирования исследуемой системы. При этом необходимо сохранять логическую структуру исследуемого процесса, учитывать последовательность протекания во времени, что позволит получить информацию о состоянии системы в заданные такты времени [6];
- с использованием статистического моделирования. То есть, с получением статистических данных о процессах в моделируемой системе.

Следует отметить, что в данной работе будут использоваться методы имитационного моделирования. Самарский А. А. и Михайлов А. П. [7] на современных этапах развития информационных технологий под имитационным моделированием понимают математическое моделирование в целом, так как с высоким темпом роста компьютерных технологий и информатизации общества строить модель без использования ЭВМ неэффективно. Помимо этого, авторы выделяют три этапа математического моделирования: модель – алгоритм – программа.

На первом этапе строится «эквивалент» объекта, представленный в виде математического выражения, отражающего важнейшие его свойства. Данная

математическая модель исследуется теоретическими методами, с целью получить априорную информацию (предварительные знания) об объекте. Конечно, не стоит забывать случаи, при которых никакой предварительной информации об объекте использовать не представляется возможным. В таких ситуациях принято говорить, что объект представлен в виде «черного ящика» и никакое математическое выражение, описывающее объект, построить невозможно. Как действовать в подобных условиях подробно описывает А. В. Медведев в своей монографии [8].

Второй этап заключается в выборе алгоритма, позволяющего реализовать модель на ЭВМ. Модель представляется в форме, удобной для реализации численных методов. Авторы отмечают, что вычислительные алгоритмы не должны искажать основные свойства модели. Выбираются методы, позволяющие найти искомые величины с заданной точностью, к примеру, описанные в данной книге [9].

Третий этап заключается в создании программы, реализующей алгоритмы, используемые для построения модели.

Таким образом, задача моделирования имеет достаточно широкое распространение. Строить модели исследуемых процессов или явлений необходимо в научной деятельности, в производстве, тяжелой промышленности, при конструировании. Ко всему прочему, существует и более широкий спектр областей, для которых актуальна данная задача: проектирование транспортных развязок, медицинская диагностика, анализ рынка валюты и котировок в экономике и много-многое другое.

Такой большой спектр задач требует большого количества методов и подходов для конструирования модели, ведь невозможно использовать лишь одно правило под все задачи одновременно. Поэтому, данное направление активно развивается, особенно с появлением более усовершенствованной вычислительной техники, способной обрабатывать большое количество данных одновременно.

1.2 Машинное обучение

Машинное обучение можно считать одним из научных направлений, позволяющих решать задачи моделирования. Этот термин комбинирует в себе большое количество алгоритмов, способных работать с данными, анализировать их, делать прогнозы, выявлять закономерности в данных и т.д., с использованием вычислительной техники.

Машинное обучение, в зависимости от количества имеющейся априорной информации, принято подразделять на два типа [10]: обучение с учителем и обучение без учителя.

В первом случае, в качестве исходных данных заданы пары «объект-ответ», так называемые прецеденты, которые образуют обучающую выборку. На основании данной обучающей выборки отыскивается функциональная зависимость полученных входов и выходов и создается алгоритм, позволяющий построить модель, которая наилучшим образом будет соответствовать исходной информации. Для того, чтобы проверить качество восстанавливаемой зависимости пользуются критериями качества, которые показывают насколько хорошо модель описывает наблюдаемые данные. Данные критерии используются в алгоритме обучения, где отыскивается такой набор параметров модели, при котором значение критерия будет оптимальным. Более подробно критерии качества построения оценок рассматриваются в данной работе [11]. К задачам, использующим обучение с учителем можно отнести задачи классификации, восстановление регрессии, прогнозирование. Более детально они будут разобраны в следующих параграфах.

Второй случай применим, когда в качестве исходных данных имеется выборка входов объекта, однако выходные значения неизвестны. В этом случае оценивается сходство исходных значений выборки между собой. Классическая задача обучения без учителя – это задача кластеризации, где исходная выборка разбивается на непересекающиеся подмножества (кластеры). Принадлежность

каждого значения определенному кластеру устанавливается путем оценивания сходства между объектами (например, оценка расстояния между объектами).

Основная и наиболее встречающаяся проблема в машинном обучении – это переобучение. С данной проблемой можно столкнуться, если достигается высокое качество работы алгоритма на обучающей выборке, однако при использовании экзаменационной выборки (то есть таких значений выборки, которые не являются элементами обучающей) качество резко и заметно падает. Подобный эффект может произойти из-за того, что в обучающей выборке обнаруживаются такие закономерности, которых не наблюдается в экзаменующей.

Более подробно о типах машинного обучения, различных алгоритмах, применимых для обучения с учителем и без, а также о разрешении проблемы переобучения можно ознакомиться в книге [12] профессора Флаха П.

Методы машинного обучения используются при решении широкого спектра задач, особенно на современных этапах развития технологий. Данные методы особенно актуальны для медицины. Так, например, в статье [13] описывается применение алгоритмов машинного обучения, позволяющих анализировать результаты магнитно-резонансной томографии с целью выявления различных психологических расстройств, что ранее было невозможно, используя только рентгенограмму. Авторы также отмечают, что благодаря машинному обучению стало возможным выставить диагноз с учетом индивидуальных особенностей пациента, а не основываясь лишь на обобщенных данных.

Помимо этого, машинное обучение широко используется в такой современной области, как «интернет вещей» (физические объекты, оснащенные технологиями с доступом в интернет, позволяющими взаимодействовать друг с другом, а также управляться человеком удаленно). В работе [14] авторы отмечают важность развития данной области в будущем, так как с ростом использования технологий растет объем информации, который необходимо обрабатывать: классифицировать, кластеризовать и т.д. Вследствие увеличения объема доступных данных, стремительно развивается такая область, как

«большие данные». Со всеми этими задачами эффективно справляются алгоритмы машинного обучения, что подробно описывается в данной работе.

1.3 Классификация. Прогнозирование

Если машинное обучение, как было описано в предыдущем параграфе, это совокупность методов, алгоритмов, подходов, позволяющих решать различные задачи анализа, моделирования данных с использованием средств вычислительной техники, то прогнозирование и классификация являются частью этого направления.

Исторически, такое понятие, как «классификация» зародилось задолго до создания первых компьютеров. К примеру, знаменитая периодическая система элементов Д. И. Менделеева, которая классифицирует химические элементы по главному признаку – заряду атомного ядра с упорядочиванием элементов внутри каждого класса. Или же иерархическая классификация растений и видов, основанная на понятии их сходства, предложенная М. Адансоном. Приведенные примеры показывают актуальность этой задачи еще с давних времен. Однако, задача классификации широко распространена и в сегодняшние дни.

Как было отмечено в предыдущем параграфе, задача классификации – это задача машинного обучения с учителем. Имеется множество объектов, которые распределены между классами. Для каждого объекта известно, к какому множеству он относится (обучающая выборка). Имеется также набор данных, для которых неизвестна принадлежность к классу (экзаменующая выборка). Требуется построить алгоритм, который будет способен классифицировать произвольный объект из экзаменующей выборки, то есть указать наименование (или номер) класса, к которому относится данный объект.

Типичным примером задачи, в которой необходимо классифицировать объекты, является задача медицинской диагностики. В качестве объектов здесь выступают пациенты. Их признаковым описанием служат результаты обследований, симптомы заболевания и применявшиеся методы лечения. При накоплении достаточного количества прецедентов (то есть формализованной

история болезни пациента) можно решать различные задачи, например, классифицировать вид заболевания, находить синдромы – наиболее характерные для данного заболевания совокупности симптомов, а также другие закономерности. Таким образом, потребность в данном типе задач достаточно высока, особенно при большом количестве прецедентов, которые человек не в состоянии обобщить и проанализировать мгновенно.

Еще одним типом задач машинного обучения с учителем является прогнозирование. Можно сказать, что прогнозирование – это частный случай классификации или задачи восстановления регрессии. В узком смысле этого термина прогнозирование понимается как определение будущих значений временного ряда по его текущим и прошлым значениям [15].

Существует большое количество методов для решения задач прогнозирования. Например, в книге [16] описывается целый комплекс методов и алгоритмов для восстановления функции регрессии по наблюдениям. Для построения прогнозов с высокой точностью, в основном необходимо обладать обширным количеством входных данных. Помимо этого, в своей книге [17] автор утверждает, что качество построения моделей прогнозирования можно увеличить с помощью конструирования самоорганизующихся систем. Цитируя профессора можно сказать, что «при моделировании по принципу самоорганизации нет незаменимых аргументов. Модель может быть построена на различных наборах аргументов, что не скажется на объективности прогноза, но при этом должна быть обеспечена достаточная свобода выбора последующих решений. Последнее исключает необходимость использования переменных, труднодоступных для измерения или не поддающихся формализации».

Активное применение данный класс задач находит в экономической сфере. Решаются такие глобальные задачи как анализ и прогноз экономической ситуации в стране для планирования бюджета, прогнозирование уровня инфляции или возникновения кризиса. Однако, с помощью методов прогнозирования можно решать и более простые задачи. Например, предсказание курса валюты, стоимости недвижимости в городе и т.д. Так, в

статье [18] автор рассматривает возможность прогнозирования российской и американской валюты с использованием нейронных сетей.

В задачах оценивания рисков модели прогнозирования также находят свое прикладное значение. Так, например, в статье [19] группа ученых сконструировала модель прогнозирования различных несчастных случаев, связанных с железнодорожным транспортом. Их модель позволяет определить какие факторы влияют на несчастные случаи больше всего, среди факторов: среднее и среднесуточное железнодорожное движение, ширина дороги, длина полотна дороги, географический регион и т.д. Оценка строится на базе нелинейного и обычного метода наименьших квадратов.

Исходя из вышеизложенного можно отметить высокую необходимость и применимость данных методов в прикладных задачах. Помимо описанных примеров, классификация и прогнозирование активно используются в других отраслях, например, промышленности или в образовании.

1.4 Параметрическое моделирование

В предыдущих параграфах обобщенно рассматривались различные методы построения математических моделей. Это было необходимо с целью осветить всю широту существующих способов моделирования, а также с целью показать прикладную значимость данного направления. Теперь же, в соответствии с тематикой дипломной работы, необходимо более углубленно изучить некоторые аспекты моделирования систем.

Одним из наиболее значимых разделов моделирования является теория идентификации систем. Идентификация – это одно из самых важных направлений в теории управления. Ее задача связана с построением модели на основании наблюдений, полученных в условии функционирования объекта по его входным и выходным переменным [20]. Профессор Д. Гроп в своей монографии [21] считает, что задачу идентификации следует рассматривать как сопряженную с задачей управления системой. Автор приводит такой пример: «Нельзя управлять системой, если она не идентифицирована либо заранее, либо

в процессе управления. Например, мы не можем управлять автомобилем, пока не познакомимся с его реакцией на поворот руля, нажатие акселератора или тормоза, то есть пока не ознакомимся со свойствами автомобиля». Таким образом, управлять системой намного эффективнее, осуществив идентификацию (или определив реакцию) этой системы.

Как правило, постановка задачи идентификации систем зависит от количества имеющейся априорной информации об объекте исследования. Так, различают задачу идентификации в «широком» и «узком» смысле. Для параметрического моделирования характерна постановка задачи в «узком» смысле, так как исследователь обладает обширным количеством априорной информации. В таких случаях известна структура изучаемого объекта, значения входов, выходов и т.д. Данные методы предполагают определение структуры модели по имеющейся априорной информации с точностью до вектора параметров на начальном этапе исследования. Затем, оценку этих параметров осуществляют известными методами [9].

Конечно, имеют место такие ситуации, когда структура исследуемого процесса на начальных этапах моделирования может быть определена неверно. В таком случае, правдивость дальнейших результатов ставится под сомнение. Данная проблема исследовалась в этих работах [22, 23].

Параметрические методы моделирования имеют высокую применимость во всех видах и отраслях науки и техники. Так, например, в работе [24] предложен метод оценивания параметров гармоник электроэнергетической системы в режиме реального времени. Данное исследование необходимо для выявления искажения сигналов (гармоник) напряжения и тока, связанных с растущим числом использования электронных устройств, вызывающих данный эффект. Гармоники могут спровоцировать помехи в цепях связи и оборудовании, увеличить потери и нагрев в электромагнитных устройствах и т. д.

Помимо этого, параметрические модели используются в медицине [25], экономике [26], что подтверждает значимость данных методов в моделировании.

1.5 Непараметрическое моделирование

Как отмечалось в предыдущем параграфе, существуют два подхода к идентификации: в «узком» и «широком» смысле. Так, параметрические методы применимы для задач в «узком» смысле. Однако, когда априорной информации слишком мало для того, чтобы построить структуру модели, пользуются непараметрическими моделями, то есть идентификацией в «широком» смысле. По сути, непараметрические методы можно противопоставить параметрическим. Такие модели реализуют определенное сглаживание экспериментальных данных, причем сама функция-оценка непараметризуема, то есть, не может быть задана в форме разложения по определенной системе координат функций. Традиционными непараметрическими оценками такого рода можно считать оценки регрессии Надарая-Ватсона и парzenовские оценки плотности вероятности.

Профессор В. Я. Катковник в своей книге [27] утверждает, что «различие между параметрическими и непараметрическими методами не носит столь уж принципиального характера». Автор предлагает обширный класс непараметрических оценок, включающий в себя, в качестве частных случаев, традиционные (параметрические) оценки, базирующиеся на методе локальной аппроксимации. Суть данного метода заключается в использовании скользящих локально-параметрических моделей, как решения специальных экстремальных задач для определения непараметрических оценок. В оценках данного типа используется параметр локальности, который определяет область применимости локально-параметрической модели. При увеличении данного параметра расширяется область применимости. Увеличив данный параметр до предела, локально-параметрическая модель превращается в обычную параметрическую. Как утверждает профессор: «такое погружение параметрических оценок в класс непараметрических в принципе означает, что потенциальные возможности последних по крайней мере не хуже возможностей параметрических оценок».

Данное утверждение оправдано с точки зрения опыта расчетов, который показывает существенную точность непараметрических оценок.

Достаточно актуальной проблемой для непараметрических оценок является выбор параметра сглаживания и доверительных интервалов. Один из основоположников данного раздела математической статистики профессор Хардле В., в своей книге [28] показывает, что все методы сглаживания в асимптотическом смысле, по существу, эквивалентны ядерному сглаживанию, то есть обосновывает решение данных проблем для того метода, который математически более удобен и проще для понимания на интуитивном уровне (ядерное сглаживание).

Непараметрические методы используются для обширного круга задач. Так, например, в работе [29] описываются результаты прогнозирования относительной мощности ветряных электрических установок в зависимости от сезонных и погодных факторов с применением непараметрической модели k -ближайших соседей. Более того, непараметрическая идентификация активно используется в космической отрасли. Автор статьи [30] разработал модель оценки показателей эффективности малорасходных вентиляторов и электронасосных агрегатов спутников связи, включая их зависимость от конструктивных параметров рабочих элементов и технологических условий эксплуатации с использованием оценки Надарая-Ватсона.

Таким образом, непараметрические методы оценивания функционалов позволяют строить модели в условиях малой априорной информации об объекте исследования, что открывает новые границы в области анализа и позволяет решать практически любые поставленные задачи.

1.6 Анализ данных. Выбросы. Робастный анализ

В предыдущих параграфах был рассмотрен достаточно широкий класс методов, позволяющих анализировать объекты или явления. Для более полного и завершенного изложения следует остановиться на заключительном пункте обзора тематики дипломной работы – это анализ данных, а также проблемы,

возникающие в данной области. Однако, чтобы приступить к изучению этой тематики следует обобщить некоторую информацию, изложенную в параграфах выше.

Так, методы обработки данных для построения моделей можно условно описать следующим образом. На начальных этапах, в понимании классической прикладной математики, объект изучался методами вычисления одних характеристик по известным значениям других характеристик данного объекта. При этом модель объекта считается известной, а зависимость между характеристиками представлена в виде уравнений, систем уравнений или неравенств. Позже, появились другие задачи анализа объектов, которые предполагали построение модели с точностью до вектора параметров (параметрическая идентификация) с использованием таблиц (или протоколов наблюдений) значений вход-выход. После, с появлением кибернетики, стали решаться задачи анализа «черного ящика» (непараметрическая идентификация), когда математическая модель, описывающая закономерности влияния характеристик не известна, однако в наличии имеется таблица экспериментальных данных «объект-признак». При таком подходе выбор модели и ее параметров делается путем проверки разных эмпирических гипотез, основываясь на доступной априорной информации. Однако, как и для задач параметрической идентификации имеется таблица данных, которую необходимо интерпретировать. Возникающий при этом круг задач и составляет направление, именуемое задачами анализа данных. То есть, анализ данных можно определить как совокупность методов и средств извлечения из некоторых данных информации для принятия решений.

Данные в таблицах могут быть представлены, как в виде чисел, так и в виде символов или логических выражений. Более научно их принято подразделять на количественные и качественные данные. Качественными (не числовыми) данными могут быть, к примеру, пол человека, его социальное положение или же материал стен какого-либо здания, тип планировки и т.д. Помимо всего прочего, для данных, представленных в таблицах, существуют различные

шкалы, по которым эти данные измеряются. Большаков в своей книге [31] для качественных данных выделяет следующие шкалы: шкала наименований (номинальная), порядковая (ординальная) и шкала гиперпорядка. Количественные данные автор представляет такими шкалами как: интервальная, отношений, разностей и абсолютная.

Подробнее о методах анализа и обработки количественных данных можно ознакомиться в этой книге [32]. Обработка логических выражений и символов (качественной информации) отражена больше в анализе знаний, чем в анализе данных, где базы знаний интерпретируются на понятный для ЭВМ язык. Подробнее о методах работы с таким типом задач, а также с понятием нечетких знаний можно ознакомиться здесь [33]. В этой работе данные представлены в виде чисел, следовательно, не будем заострять внимание на вышеприведенных случаях.

В монографии [34] автор утверждает, что анализ данных включает в себя решение задач двух связанных между собой направлений:

- обнаружение закономерных связей между элементами таблицы
- использование обнаруженных закономерностей для предсказания (прогнозирования) значений одних элементов таблицы по известным значениям других ее элементов.

Однако, на практике чаще всего приходится сталкиваться с ситуациями, когда таблицы с данными имеют какие-либо дефекты. Например, пропуски или выбросы.

Пропуски могут возникать по разным причинам, тем не менее, в задачах управления и идентификации это может быть связано с нарушением дискретности контроля входных-выходных переменных. Например, измерения одних переменных проводятся электрическим способом, а других – путем физико-механических испытаний или лабораторных анализов. Если количество пропусков в данных и их влияние на общий объем информации незначительное, то пустые значения (строки) можно исключить. В других случаях пользуются различными методами, которые, в частности, описаны в этой работе [35].

Наличие выбросов в исходной выборке может быть обусловлено какими-либо ошибками оператора (человеческий фактор) в момент записи данных или же неисправностями контролирующих приборов, а также неточной математической моделью. Значение математического ожидания, дисперсии, коэффициента корреляции, критерия наименьших квадратов для настройки регрессионных моделей, все эти вещи сильно подвержены влиянию даже хотя бы одного случайного выброса. Значения этих параметров будут сильно искажены, а результаты моделирования и других математических экспериментов будут не точными.

В связи с возникновением данной проблемы, применяются специальные методы, которые позволяют исключать или не учитывать аномальные измерения выборки. Такие методы принято называть робастными (от английского «robust»: здоровый, крепкий, сильный). На сегодняшний день существует целая самостоятельная ветвь математической статистики – робастная статистика (или робастный анализ), которая описывает большой спектр алгоритмов и методов поиска, а также исключения выбросов из данных с целью повышения точности конструируемых моделей. Особенно активно данное направление развивается за рубежом. Так, с подробным описанием методов, включающих примеры из реальной жизни можно ознакомиться в работах английских [36] и канадских [37] ученых. Однако и в России существует достаточно современная литература, посвященная систематизации знаний в робастной статистике, а также описанию алгоритмов, например, книга [38], в которой автор систематизирует известные знания о робастном оценивании, а также приводит модификации и обобщения некоторых алгоритмов (Ходжеса-Лемана, средней разности Джини, интерквартильного размаха и др.).

Применение робастных алгоритмов является значимым, если исследователь стремится достигнуть высокой точности результатов моделирования. Например, для медицинской отрасли точность является одним из наиболее важных факторов. В статье [39] описан широкий класс задач, решаемых с помощью методов робастной статистики. С помощью

разработанных компьютерных систем повышается точность диагностики различных заболеваний, решаются задачи управления банковскими рисками, в частности, эконометрическое прогнозирование и многое другое. Помимо этого, развиваются и совершенствуются уже известные методы робастного оценивания [40].

Однако от выбросов необязательно всегда избавляться. В. И. Тихонов в своей монографии [41] приводит практические примеры, основанные на изучении характеристик и свойств выбросов. Так, например, это находит свое применение в строительной механике для расчетов механических конструкций на прочность. В данном случае, параметры самой конструкции и внешние нагрузки носят случайный характер. Цель расчета заключается в получении гарантии того, что за время эксплуатации не наступит ни одно из недопустимых предельных состояний. Помимо этого, выбросы измеряются в большинстве сейсмических или медицинских приборах, в частности, медицинские приборы регистрируют биотоки сердца и мозга, базирующиеся на измерении высоты и длительности выбросов электроосциллограмм, а также интервалы между выбросами. Некоторые резко отличающиеся измерения, полученные при помощи электроэнцефалограмм, могут свидетельствовать о патологиях здоровья человека.

Таким образом, был рассмотрен последний класс задач, связанный с анализом данных и построением моделей систем или явлений. Конечно, существует большое количество других направлений в исследуемой области анализа, однако они в меньшей степени относятся к рассматриваемой тематике дипломной работы.

1.7 Выводы по главе 1

В главе было отмечено, что построение моделей исследуемых процессов или явлений имеет широкое применение для различного класса задач. Была рассмотрена классификация типов моделирования систем. В итоге обобщенно структурную схему моделирования можно представить следующим образом. На

первом этапе выбирается тип моделирования (физическое, математическое), после чего определяется алгоритм или метод, позволяющий построить модель. Выбор основывается на количестве и качестве априорной информации, а также от требований, поставленных перед аналитиком. Если выборка с дефектами (выбросы, пропуски), то решаются вопросы по их устранению.

Из всей совокупности рассмотренных тем стоит выделить непараметрическую идентификацию систем, прогнозирование и робастную регрессию, так как в следующих параграфах будут рассмотрены методы, относящиеся именно к ним. Выбор данных тематик был сделан на основании актуальности упомянутых областей науки, а также высокого уровня применимости этих направлений на практике, ведь, довольно часто приходится сталкиваться с дефицитом априорной информации, а также с различными дефектами выборки, от которых следует избавляться с целью повышения точности конструируемой модели. Помимо этого, достаточно интересна задача прогнозирования данных, в связи с возросшей потребностью исследований в данной области, а также в связи с очевидным желанием обладать информацией, которая с какой-то долей вероятности будет существовать в будущем.

2 Алгоритмы обработки данных с выбросами

Как уже отмечалось в предыдущей главе, методы робастного оценивания имеют высокий уровень значимости при работе с выборкой наблюдений, содержащей шум или выбросы. Однако более подробно данный вопрос еще разобран не был. Какие методы существуют? В чем их принципиальное различие? Далее, об этом и пойдет речь.

2.1 Классификация методов работы с выбросами

Разработка методов обработки данных с выбросами началась ближе к концу двадцатого столетия. Проанализировав литературу по данной тематике можно условно классифицировать предложенные методы на следующие группы:

– робастные алгоритмы обработки выборки с выбросами. Такие методы еще называют «robust learners». Они меньше подвержены влиянию выбросов в выборке наблюдений. Задачи данного типа рассматриваются в работах Хьюбера [36]. В качестве примера можно выделить алгоритм RIPPER, предложенный Вильямом Кохеном [42], как модификация его же алгоритма IREP. Класс таких задач реализует правило пропозиционального (относящегося к высказываниям или предложениям) обучения. Он основан на ассоциативных правилах с REP (reduced error pruning – сокращение обрезки ошибок). Данная методика была достаточно распространенной и эффективной для алгоритмов, построенных с помощью дерева решений. Для REP исходная выборка разбивается на два набора: «растущий» набор и набор «подрезки». На первом этапе формируется «растущий» набор правил, с использованием какого-либо эвристического правила. Данный избыточный набор правил затем многократно упрощается, с помощью одного из множества операторов «обрезки». Таким оператором, к примеру, может служить удаление любого единственного условия или любого единственного правила. На каждом этапе упрощения выбирается такой оператор «обрезки», который дает наибольшее уменьшение ошибки на наборе «обрезки».

Упрощение заканчивается, когда применение любого оператора «обрезки» приводит к увеличению ошибки в наборе обрезки.

– методы предварительной обработки данных (цензурирование данных) – исключение из рассмотрения подозрительные на выбросы точки. Например, модификация алгоритма С4.5, представленная в работе [43]. В статье авторы описывают свою работу над повышением качества обучающей выборки с помощью обнаружения и исключения ошибочно классифицированных значений перед применением обучающих алгоритмов, тем самым увеличивая точность классификации. Значения могут быть неверно отнесены к тому или иному классу по разным причинам: ошибки в исходной выборке, неверная априорная информация о классах и т.д. Суть метода заключается в фильтровании обучающей выборки перед ее классификацией. Как утверждают авторы, их подход предполагает, что ошибки, возникающие в ходе классификации, никак не зависят от модели, с помощью которой данная выборка была получена.

2.2 Робастная оценка регрессии №1

Приведем подробное описание робастного алгоритма для непараметрической оценки регрессии, который был предложен А. И. Рубаном [44]. Постановка задачи в данном случае выглядит следующим образом. Предположим, что рассматривается дискретно-непрерывный процесс, общепринятая схема которого приведена на рис. 1, где в качестве обозначений принято: A – неизвестный оператор объекта; $y(t) \in \Omega(y) \subset R^1$ – выходная переменная процесса; $u(t) = (u_i(t), i=1, \dots, m) \in \Omega(u) \subset R^m$ – векторное входное воздействие; $\zeta(t)$ – векторное случайное воздействие; t – непрерывное время; H^u , H^y – каналы связи; $h^u(t)$, $h^y(t)$ – случайные помехи измерений; $\{u_i, y_i, i=1, \dots, s\}$ – обучающая выборка, где s – объем выборки.

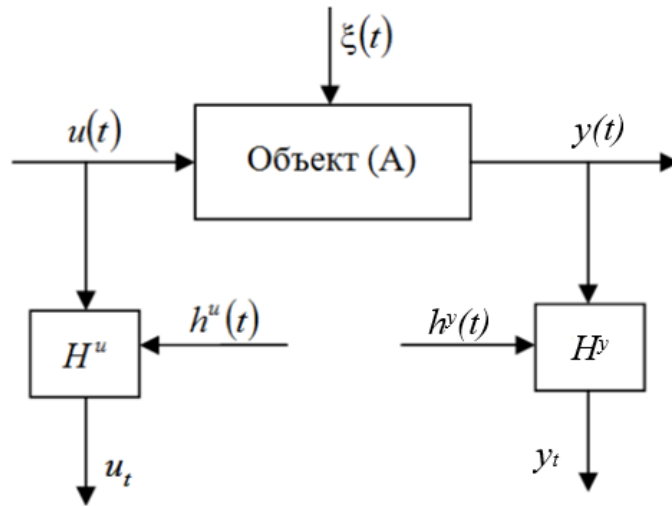


Рисунок 2 – Схема исследуемого процесса

В качестве непараметрической оценки регрессии рассматривается одномерная непараметрическая оценка Надарая-Ватсона, математическое описание которой в общем виде выглядит следующим образом:

$$\eta_n = \frac{\sum_{i=1}^s y_i \Phi\left(\frac{u - u_i}{c_s}\right)}{\sum_{i=1}^s \Phi\left(\frac{u - u_i}{c_s}\right)}, \quad (1)$$

где c_s – параметр размытости ядра;

$\Phi(\bullet)$ – ядерная функция, которые удовлетворяют условиям сходимости [45].

Оценка (1), для каждого фиксированного u , удовлетворяет квадратичному критерию с весами $\Phi(\bullet)$:

$$I_1(u) = \sum_{i=1}^s (y_i - \eta_n)^2 \Phi\left(\frac{u - u_i}{c_s}\right) = \min_{\eta_n}. \quad (2)$$

Робастный аналог оценки (1) отыскивается из минимума модульного взвешенного критерия:

$$I_2(u) = \sum_{i=1}^s |y_i - \eta_{2n}| \Phi\left(\frac{u - u_i}{c_s}\right) = \min_{\eta_{2n}}, \quad (3)$$

где итерационная формула для расчета устойчивой к выбросам оценки η_{2n} имеет вид:

$$\eta_2^{l+1} = \sum_{i=1}^s \left[\frac{|y_i - \eta_{2n}^l|^{-1} \Phi\left(\frac{u - u_i}{c_s}\right)}{\sum_{j=1}^s |y_j - \eta_{2n}^l|^{-1} \Phi\left(\frac{u - u_j}{c_s}\right)} \right] y_i \equiv \sum_{i=1}^s \omega_2^l \left(\frac{u - u_i}{c_s} \right) y_i, \quad (4)$$

где $l=0,1,2,\dots$

Итерационная коррекция оценки заканчивается, как только выполняется следующее условие:

$$|\eta_2^{l+1} - \eta_2^l| \leq \varepsilon, \quad (5)$$

где ε – настраиваемый параметр.

При выполнении условия (5) η_2^{l+1} дает минимум функции качества $I_2(u)$, то есть $\eta_{2n} = \eta_2^{l+1}$.

Наличие выброса среди значений выхода объекта y_i в какой-либо точке u_i , $i=1,\dots,s$ приводит к уменьшению значения весового коэффициента $\omega_2(\bullet)$, что в конечном счете позволяет нивелировать влияние выброса на оценку η_{2n} .

2.3 Робастная оценка регрессии №2

Еще один тип оценок, настраиваемых в ходе выполнения алгоритма, приведен в работе [46] Кирик Е. С. Постановка задачи аналогична, описанной в параграфе 2.2. За основу берется классическая непараметрическая оценка Надарая-Ватсона (1), однако в данном алгоритме рассматривается ее многомерная интерпретация:

$$\hat{y}_1(c_s) = \frac{\sum_{i=1}^s y_i \prod_{j=1}^m \Phi\left(\frac{u_j - u_{ji}}{c_s}\right)}{\sum_{i=1}^s \prod_{j=1}^m \Phi\left(\frac{u_j - u_{ji}}{c_s}\right)}, \quad (6)$$

где m – количество входов объекта.

Оптимальный параметр размытости c_s настраивается с использованием скользящего экзамена исходя из условия минимума квадратичного критерия:

$$\omega = \sum_{i=1}^s (y_i - \hat{y}_{li})^2 \rightarrow \min_{c_s}. \quad (7)$$

В связи с тем, что оценка (6) точечная взвешенная, очевидно, она является чувствительной к влиянию выбросов в данных. Для того, чтобы перейти к робастному аналогу оценки (6), в данном алгоритме предлагается исследовать невязки $\varepsilon_i = y_i - \hat{y}_{li}(c_s)$, $i = 1, \dots, s$. Упорядочив величины ε_i , $i = 1, \dots, s$ по возрастанию, получим вариационный ряд:

$$\varepsilon^1 \leq \varepsilon^2 \leq \dots \leq \varepsilon^s, \quad (8)$$

где $\varepsilon^1 = \min(\varepsilon_i)$;

$$\varepsilon^s = \max(\varepsilon_i), i = 1, \dots, s.$$

На первом этапе необходимо восстановить функцию плотности невязок ε_i , $i = 1, \dots, s$. Для этого воспользуемся общим принципом построения оценки регрессии (1) и в результате получим:

$$p_n(\varepsilon, C_n^\varepsilon) = \frac{1}{s C_n^\varepsilon} \sum_{i=1}^s \Phi\left(\frac{\varepsilon - \varepsilon_i}{C_n^\varepsilon}\right), \quad (9)$$

где C_n^ε – оптимальный параметр размытости ядра оценки (9), который определяет качество «очистки» выборки от выбросов. Область определения (9) задается минимальным и максимальным значениями вариационного ряда, составленного из значений невязок (8).

Для того, чтобы определить «очищенную» выборку наблюдений, исследуем (9) на ближайший слева и справа к нулю минимумы. Элементы, значения невязок которых лежат между вышеупомянутыми минимумами, будут составлять «очищенную» выборку, остальные – выбросы.

Таким образом, робастная оценка регрессии примет вид:

$$\hat{y}_2(C_n, C_n^\varepsilon) = \frac{\sum_{i=1}^s y_i \prod_{j=1}^k \Phi\left(\frac{u_j - u_{ji}}{C_n}\right) I(\varepsilon_i, C_n^\varepsilon)}{\sum_{i=1}^s \prod_{j=1}^k \Phi\left(\frac{u_j - u_{ji}}{C_n}\right) I(\varepsilon_i, C_n^\varepsilon)}, \quad (10)$$

где $I(\bullet)$ – индикаторная функция:

$$I(\varepsilon_i, C_n^\varepsilon) = \begin{cases} 0, & \varepsilon_i \in [\varepsilon^1, \varepsilon^{-0}(C_n^\varepsilon)) \cup (\varepsilon^{+0}(C_n^\varepsilon), \varepsilon^s], \\ 1, & \varepsilon_i \in [\varepsilon^{-0}(C_n^\varepsilon), \varepsilon^{+0}(C_n^\varepsilon)], \end{cases} \quad (11)$$

где $\varepsilon^{-0}(C_n^\varepsilon)$ – наибольший отрицательный нуль функции (9);

$\varepsilon^{+0}(C_n^\varepsilon)$ – наименьший положительный нуль функции (9).

Параметр размытости C_n^ε , который был упомянут выше, определяется из условия минимума следующего критерия:

$$W_{p_n} = \frac{1}{S} \sum_{i=1}^s (y_i - \hat{y}_{2i}(C_n, C_n^\varepsilon)) \rightarrow \min_{C_n^\varepsilon}. \quad (12)$$

При известном оптимальном параметре размытости C_n^ε поиск оптимального параметра C_n для «очищенной» выборки состоит в минимизации следующего функционала:

$$W_{\hat{y}_2}^2(C_n) = \frac{1}{s} \sum_{i=1}^s (y_i - \hat{y}_{2i}(C_n, C_n^\varepsilon))^2 I(\varepsilon_i, C_n^\varepsilon) \rightarrow \min_{C_n}. \quad (13)$$

Тем не менее, даже при отсутствии выбросов, критерий (12) остается работоспособным. В таком случае оптимальным C_n^ε является такой параметр, при котором в категорию выбросов не попадает ни один элемент. Данное свойство алгоритма робастного оценивания показывает, что предложенный подход устойчив к ошибочной априорной информации о наличии или отсутствии выбросов в обучающей выборке.

2.4 Непараметрический алгоритм цензурирования данных с выбросами

Ранее, в параграфах 2.2 и 2.3 был описан ряд робастных алгоритмов, при которых влияние выбросов снижается путем нахождения оценки заданной структуры из условия минимума некоторого функционала. Однако, в параграфе 2.1 было упомянуто, что существует также и другой подход – цензурирование данных. В этом параграфе будет рассмотрен алгоритм цензурирования данных, описанный в работе [47].

Постановка задачи аналогична описанной в параграфе 2.2. На первом этапе с использованием некоторой выборки наблюдений $\{u_{ji}, y_i, i=1, \dots, s; j=1, \dots, m\}$ строится непараметрическая модель (6). Далее, для всех точек выборки проверяется следующее условие:

$$|y_i - \hat{y}_{li}| > \mu \cdot \Delta \quad i = 1, \dots, s, \quad (14)$$

где y_i – значения выхода объекта;

\hat{y}_{li} – значения выхода модели (6);

μ – настраиваемый параметр, а значение Δ определяется как:

$$\Delta = \frac{1}{s} \sum_{i=1}^s |y_i - \hat{y}_{li}|. \quad (15)$$

Алгоритм применяется для каждого значения из исходной выборки наблюдений. Если неравенство (14) выполняется, данный элемент выборки можно считать выбросом. Далее, проверив всю выборку наблюдений и запомнив номер тех точек, для которых (14) справедливо, удалим их из данных и заново построим непараметрическую оценку, но уже с другим объемом выборки s_2 . Таким образом модель для цензурированной выборки наблюдений $\{u_{ji}, y_i, i=1, \dots, s_2; j=1, \dots, m\}$ рассчитывается следующим образом:

$$\hat{y}_3(u) = \frac{\sum_{i=1}^{s_2} y_i \prod_{j=1}^m \Phi\left(\frac{u_j - u_{ji}}{c_s}\right)}{\sum_{i=1}^{s_2} \prod_{j=1}^m \Phi\left(\frac{u_j - u_{ji}}{c_s}\right)}. \quad (16)$$

Следует также добавить, что обнаружение выброса зависит от настроенного параметра μ . Результат наиболее точного выявления аномального значения будет зависеть от того, каким образом был настроен этот параметр.

2.5 Ремонт данных выборки наблюдений с выбросами

Исключение выбросов из выборки наблюдений увеличивает точность аппроксимации объекта, однако, в процессе «робастизации» модели или же цензурирования данных точки, являющиеся выбросами, удаляются, что уменьшает объем рабочей выборки. Как известно, с уменьшением объема

выборки точность моделирования снижается. Чтобы предотвратить такую ситуацию, проводится ремонт данных.

Под ремонтом данных понимают идентификацию и последующую замену грубых измерений (выбросов) значениями робастной модели.

В случае с алгоритмом из параграфа 2.3 ремонт данных производится следующим образом. После настройки робастной оценки (10), то есть нахождения оптимальных параметров C_n и C_n^ε , выбросы в выборке выхода объекта (или те значения, индикаторная функция которых равна нулю $I(\bullet) = 0$) заменяют их оценками (10) и заново рассчитывают непараметрическую модель по формуле (6).

Ремонт данных для алгоритма цензурирования выборки наблюдений, описанного в параграфе 2.4, производится на основании формулы (6). На место элементов подозрительных на выброс ставится их значение непараметрической оценки после чего заново рассчитывается модель по формуле (6).

2.6 Выводы по главе 2

Во второй главе были подробно представлены алгоритмы, позволяющие избавляться от нежелательных, резко отличающихся элементов в выборке наблюдений. Описан подробный математический аппарат реализации данных алгоритмов. Помимо этого, предложены варианты по ремонту (восстановлению) значений выхода объекта после того, как была исключена часть точек, охарактеризованная как выбросы в данных.

3 Вычислительные эксперименты

В рамках данной выпускной квалификационной работы производится сравнение двух алгоритмов (их описание приведено в параграфах 2.3 и 2.4) непараметрического восстановления функции регрессии по выборке наблюдений, содержащей выбросы. Основная цель сравнительного анализа заключается в определении лучшего алгоритма «робастизации» с точки зрения наиболее высокого значения точности конструируемой модели.

3.1 Вычислительный эксперимент с использованием робастного алгоритма №2

Предположим, что в рамках вычислительного эксперимента структура объекта задана следующим выражением:

$$y = \alpha_1 \sin(u_1) + \xi, \quad (17)$$

где α_1 – коэффициент исследуемого объекта;

ξ – центрированная помеха, генерируемая следующим образом:

$$\xi = y \cdot c \cdot k, \quad (18)$$

где c – нормально распределенная случайная величина в интервале $[-1;1]$;

k – процент помехи.

Переменная входа u_1 генерируется случайным образом по нормальному закону распределения в интервале $[1;15]$.

Математическое описание не робастной непараметрической оценки для объекта (17) было описано ранее в параграфе 2.3 – формула (6). В данном вычислительном эксперименте используется частный случай – одномерный вход u_1 . В качестве функции ядра $\Phi(\bullet)$ используется параболическая

колоколообразная ядерная функция, структура которой выглядит следующим образом:

$$\Phi(z) = \begin{cases} (0.75 \cdot (1 - |z|)^2, & |z| < 1, \\ 0, & |z| \geq 1, \end{cases} \quad (19)$$

где $z = \frac{u_1 - u_{1i}}{c_s}$.

Параметр размытости настраивается методом скользящего экзамена на основании минимума квадратичного критерия:

$$F = \frac{1}{s} \sum_{i=1}^s (y_i - \hat{y}_{1i})^2 \rightarrow \min_{c_s}. \quad (20)$$

Робастный аналог оценки (6) будет выглядеть, как частный случай формулы (10) параграфа 2.3, с одномерным входом u_1 . Ядро $\Phi(\bullet)$ в данном случае также имеет параболический вид (19).

Точность конструируемой не робастной модели (6) рассчитывается с помощью относительной ошибки аппроксимации по следующей формуле:

$$W_1 = \sqrt{\frac{\frac{1}{s} \sum_{i=1}^s (y_i - \hat{y}_{1i})^2}{\frac{1}{s-1} \sum_{i=1}^s (\hat{m}_y - y_{1i})^2}}, \quad (21)$$

где \hat{m}_y – оценка математического ожидания.

Для робастной модели (10) расчет относительной ошибки аппроксимации производится с использованием индикаторной функции (11):

$$W_2 = \sqrt{\frac{\frac{1}{s} \sum_{i=1}^s (y_i - \hat{y}_{2i})^2 I(\varepsilon_i, C_n^\varepsilon)}{\frac{1}{s-1} \sum_{i=1}^s (\hat{m}_y - y_{2i})^2 I(\varepsilon_i, C_n^\varepsilon)}}. \quad (22)$$

На первом этапе вычислительного эксперимента сгенерируем выборку наблюдений $\{y_i, u_{1i}, i=1, \dots, s\}$. Коэффициент объекта $\alpha_1=2$.

Для того, чтобы иметь возможность продемонстрировать работу алгоритма, необходимо иметь выборку, содержащую в себе выбросы. Однако, обладать такими данными не представляется возможным. Следовательно, необходимо сгенерировать их самостоятельно в рамках вычислительного эксперимента. Для этого пусть самый минимальный и самый максимальный элементы выборки умножаются на число -0.5. Данное условие позволит создать выбросы, находящиеся в области определения значений выхода объекта.

После того, как была определена выборка наблюдений с выбросами оценим ее с использованием модели (6), а также рассчитаем значение ошибки (21). Объем выборки примем $s=100$, уровень помех 5%. Модель, устойчивая к выбросам рассчитывается по формуле (10), а ее точность определяется по формуле (22). Полученные результаты моделирования в виде графических зависимостей выходов объекта y , обычной \hat{y}_1 и робастной \hat{y}_2 моделей от входа объекта u_1 представлены на рисунке ниже:

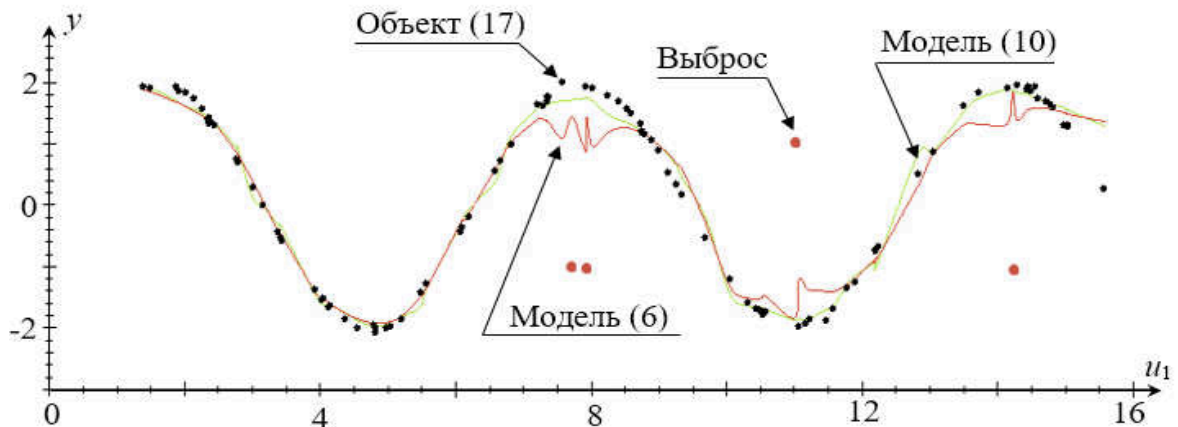


Рисунок 3 – Графическая интерпретация результатов с использованием робастного алгоритма для одномерного объекта

Отсюда и далее в текущем и следующем параграфе на графиках приняты следующие обозначения:

- зеленая цвет – значения модели (10);
- красная цвет – значения модели (6);
- черные точки – значения выхода объекта;
- красные точки – значения выбросов.

На рисунке 3 видно, что робастная модель (10) не поддается влиянию выбросов и точнее аппроксимирует выход объекта, чем модель (6), которая в свою очередь «тянется» к значению выброса, что ухудшает конечную точность моделирования. Данное утверждение подтверждают также значения относительных ошибок аппроксимации: (21) для не робастной модели $W_1=42.5\%$, (22) для робастной модели $W_2=10.1\%$. Исходя из этих значений можно судить о том, что точность моделирования с использованием непараметрической оценки устойчивой к выбросам, увеличилась примерно в четыре раза.

На практике, чаще всего, структура объекта представлена более сложным выражением. Поэтому, на втором этапе вычислительного эксперимента усложним объект (17) добавив к нему еще одну входную переменную, а также поменяем математическую структуру на выражение типа гиперболический параболоид. Таким образом получим:

$$y = \frac{u_1^2}{\alpha_1^2} + \frac{u_2^2}{\alpha_2^2} + \xi, \quad (23)$$

где α_1, α_2 – коэффициенты исследуемого объекта;

u_1, u_2 – входные переменные, генерируемые случайным образом по нормальному закону распределения в интервале $[-2;2]$.

Непараметрическая модель (6) и робастная (10), а также значения их относительных ошибок аппроксимации (21) и (22) рассчитываются аналогичным образом с учетом дополнительной входной переменной объекта (23) – u_2 .

Для вычислительного эксперимента с модифицированным объектом (23) примем объем выборки $s=300$, а значение уровня помех $k=5\%$, коэффициенты объекта равны $\alpha_1=\alpha_2=2$. Выбросы в выборке наблюдений создаются аналогично первому вычислительному эксперименту. Рассчитав значения выхода объекта, модели (6) и ее робастный аналог (10) получим следующие результаты:

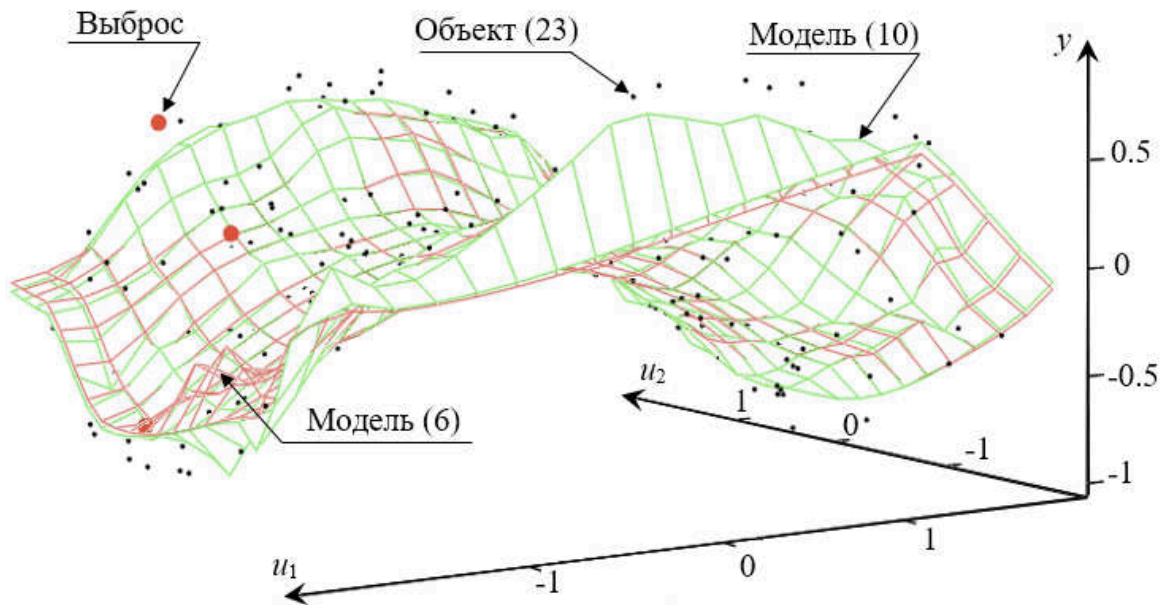


Рисунок 4 – Графическая интерпретация результатов с использованием робастного алгоритма для двумерного объекта (ракурс 1)

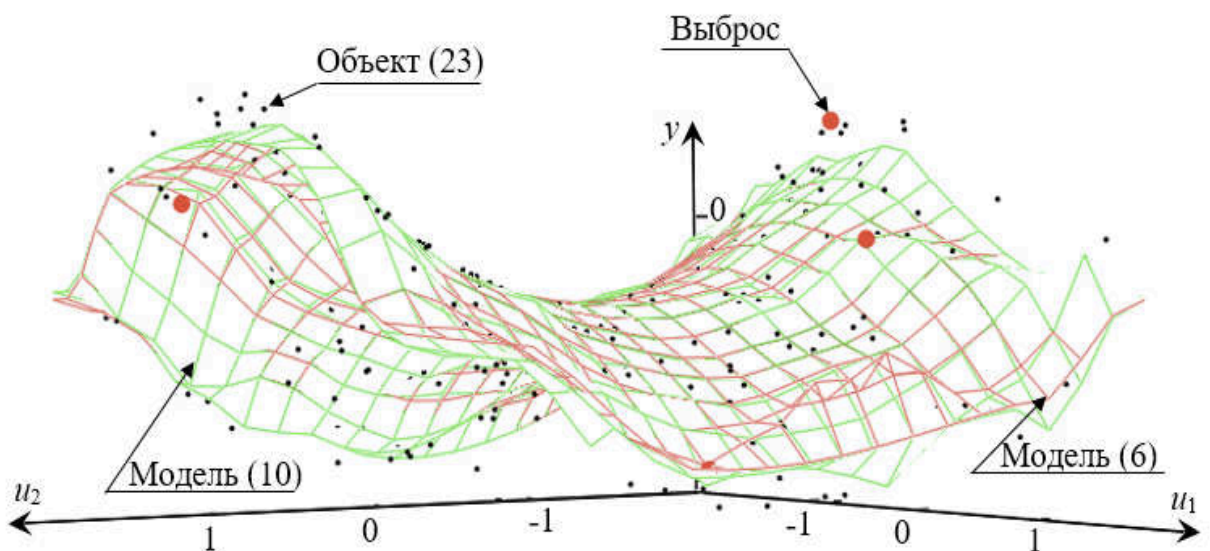


Рисунок 5 – Графическая интерпретация результатов с использованием робастного алгоритма для двумерного объекта (ракурс 2)

Точность восстановления для не робастной модели равна $W_1=30.3\%$, а с использованием робастной функции точность аппроксимации объекта равна $W_2=16\%$. Точность увеличилась, что также видно по графикам поверхностей, изображенных на рисунке 4 и 5.

3.2 Вычислительный эксперимент с использованием алгоритма цензурирования данных

Постановка задачи аналогична описанной в параграфе 3.1. В качестве объекта используется выражение (17), модель рассчитывается по формуле (6). Выборка наблюдений $\{y_i, u_{1i}, i=1, \dots, s\}$ с выбросами, ее объем $s=100$, уровень помех 5% – все эти данные аналогичны данным параграфа 3.1.

На первом этапе рассчитывается непараметрическая оценка (6). После чего с использованием формулы (14) и (15) осуществляется поиск выбросов в данных. При этом параметр μ настраивается эвристически и для данного вычислительного эксперимента равен $\mu = 4$.

После того, как выбросы были исключены из выборки наблюдений получим новое выражение для расчета модели с измененным (уменьшенным) объемом выборки s_2 , математическое описание которого было представлено в параграфе 2.4 – формула (16).

Расчет точности конструирования модели по цензурированной выборке наблюдений производится с помощью относительной ошибки аппроксимации по следующей формуле:

$$W_3 = \sqrt{\frac{\frac{1}{s_2} \sum_{i=1}^{s_2} (y_i - \hat{y}_{3i})^2}{\frac{1}{s_2 - 1} \sum_{i=1}^{s_2} (\hat{m}_y - y_{3i})^2}}, \quad (24)$$

Результаты моделирования непараметрической оценки по формуле (6), а также расчет значений непараметрической модели (16) после цензурирования выборки с использованием (14), (15) представлены ниже:

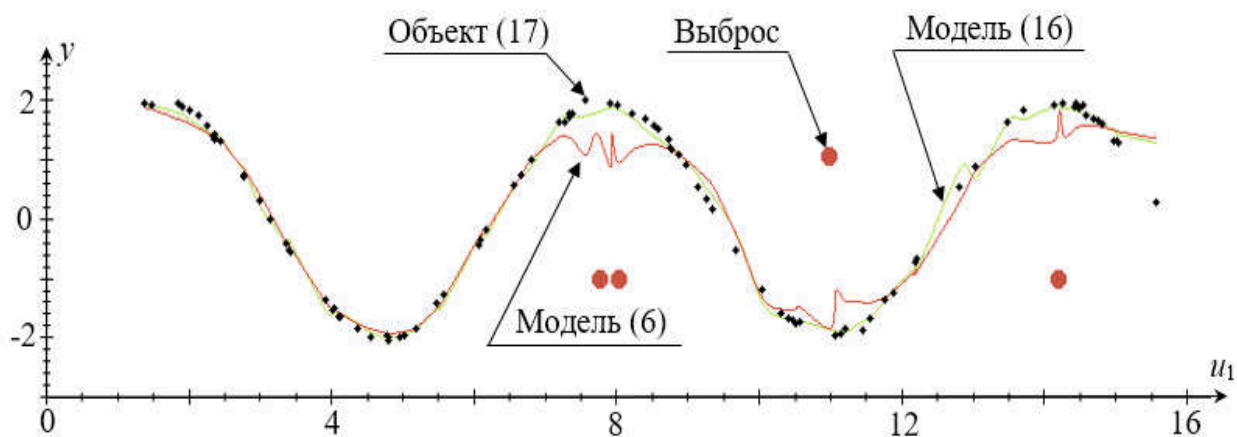


Рисунок 6 – Графическая интерпретация результатов после цензурирования данных для одномерного объекта

Значения относительных ошибок: для модели с выбросами аналогично вычислительному эксперименту параграфа 3.1 – $W_1=42.5\%$; для модели, построенной по выборке наблюдений без выбросов – $W_3=10\%$. Анализируя полученные значения, а также графики, представленные на рисунке 6, можно заключить, что алгоритм при адекватной настройке параметра μ точно справился со своей задачей, выявив и исключив все выбросы в данных.

При использовании более усложненного, двумерного объекта (23) с выборкой из предыдущего вычислительного эксперимента также проведем цензурирование данных. После «чистки» новой выборки наблюдений получим результаты, представленные на рисунке 7 и 8 ниже:

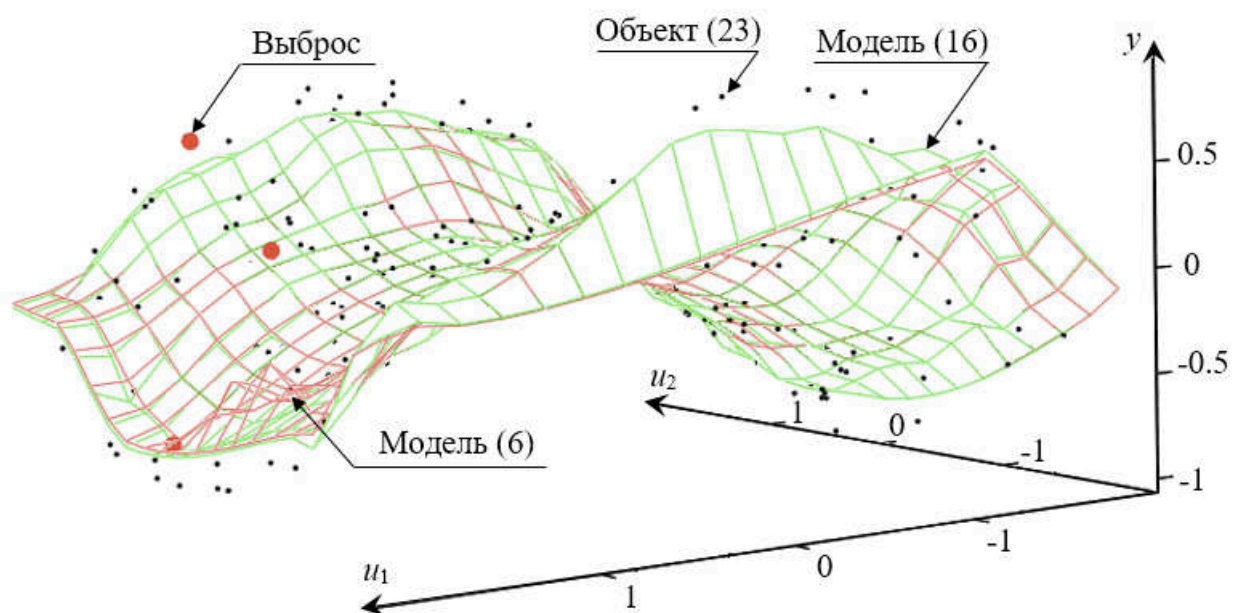


Рисунок 7 – Графическая интерпретация результатов после цензурирования данных для двумерного объекта (ракурс 1)

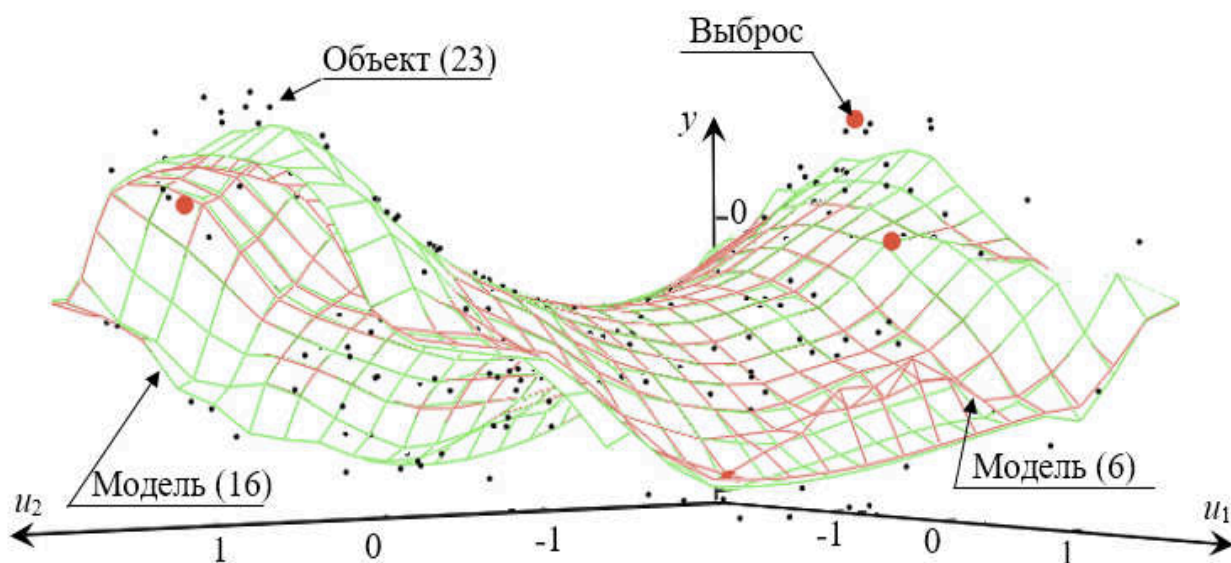


Рисунок 8 – Графическая интерпретация результатов после цензурирования данных для двумерного объекта (ракурс 2)

Относительная ошибка аппроксимации (24) в данном вычислительном эксперименте равна $W_3=15.5\%$, а параметр μ был подобран эвристически и равен $\mu=5$.

3.3 Вычислительный эксперимент по ремонту данных

Основная цель ремонта данных и описание алгоритма были разобраны в параграфе 2.5. Восстановление данных производится, как для выборки выхода объекта с одним входом из вычислительного эксперимента с робастным алгоритмом (параграф 3.1), так и для цензурированных данных (параграф 3.2).

Ремонт для выборки наблюдений, очищенной от выбросов с помощью алгоритма параграфа 3.1, будет производиться не только с помощью оценки (10), как было описано в 2.5, но и с использованием (6). Непараметрическую оценку (6) можно также использовать для ремонта точек выборки, ведь данная оценка в точке выброса «не тянется» за его значением, так как оно не попадает под «колокол» ядерной функции в процессе моделирования.

Восстановив данные с помощью робастной оценки (10) получим следующие результаты:

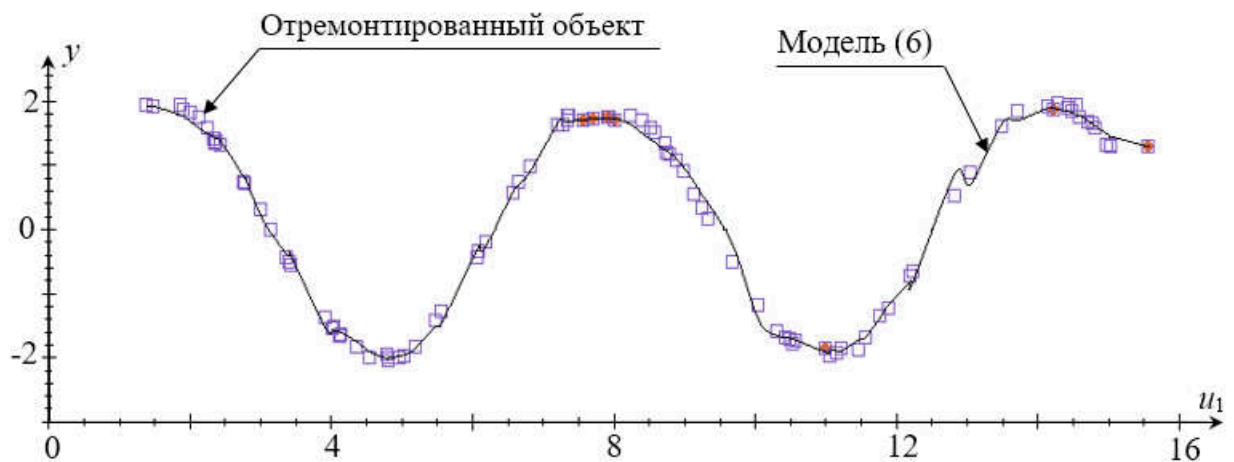


Рисунок 9 – Ремонт данных для выборки наблюдений очищенной от выбросов с помощью робастного алгоритма по (10)

Отсюда и далее в текущем параграфе, для графиков приняты следующие обозначения:

- восстановленные точки объекта – фиолетовые квадраты;
- красная точка в фиолетовом квадрате – значения, которые были отремонтированы с помощью модели;

– черная линия – значения модели (6).

Результаты вычислительного эксперимента (рисунок 9) показывают, что помимо точек, которые являлись выбросами, были восстановлены также некоторые значения «на концах» отрезка области определения выхода объекта. Такое поведение вполне характерно для оценок непараметрического типа. В связи с тем, что точность аппроксимации «на концах» объекта значительно хуже, чем на всей остальной его части, во время ремонта выборки некоторые значения могут быть также определены как выбросы в данных из-за большого значения невязки (значение индикаторной функции (11) для них равно нулю).

Далее, применяя иной подход к ремонту данных, а именно, с использованием непараметрической оценки (6), получим результаты, отображенные в виде графической зависимости, представленной на рисунке ниже:

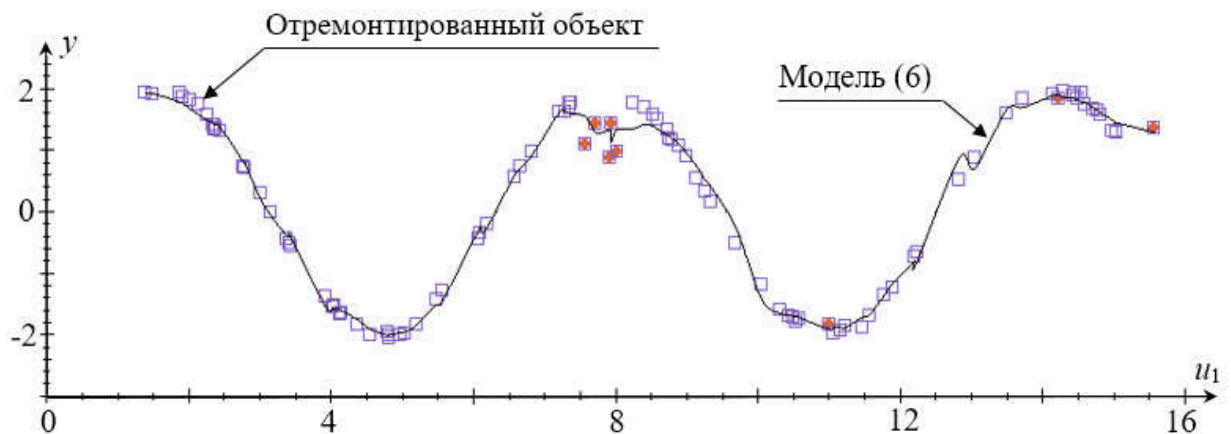


Рисунок 10 – Ремонт данных для выборки наблюдений очищенной от выбросов с помощью робастного алгоритма по (6)

На рисунке 10 видно, что для промежутка, на котором в ходе вычислительного эксперимента было сгенерировано два выброса, точность ремонта данных низкая. Не робастная непараметрическая оценка (6) для тех интервалов, в которых содержится два и более выброса, помимо «хороших» значений модели для выбросов, «притягивает» также и их «плохих» соседей, что

значительно ухудшает точность восстановления данных и портит итоговые результаты моделирования.

На следующем этапе рассмотрим ремонт данных по цензурированный выборке наблюдений. В данном случае восстановление будет производиться только по значениям модели (6).

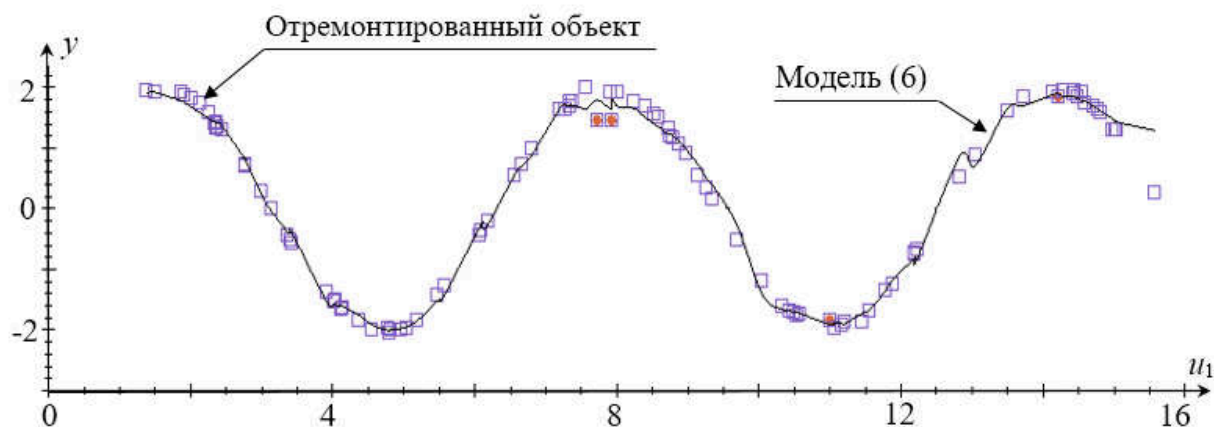


Рисунок 11 – Ремонт данных для цензурированной выборки наблюдений

Анализируя результаты, приведенные на рисунке 11, можно сказать, что заменены были только те значения, что являются выбросами. В предыдущем же случае, индикаторная функция помимо самих значений выбросов помечала точки на концах интервала, а также точки «плохого» качества, что лежали по соседству с выбросами (случай восстановления по модели (6)).

Значения относительных ошибок моделирования W , рассчитанных с учетом новых значений отремонтрованного объекта по аналогичным формулам (21, 22, 24) из предыдущего эксперимента, представлены ниже:

ошибка для робастной модели отремонтрованной по (10)	$W = 7.1\%$
ошибка для робастной модели отремонтрованной по (6)	$W = 9.8\%$
ошибка для цензурированной модели отремонтрованной по (6)	$W = 11.1\%$

3.4 Сравнение результатов работы алгоритмов

Для того, чтобы результаты сравнительного анализа были адекватными, необходимо осуществить моделирование объекта по формулам (6), (10), (16)

несколько раз. Провести такой эксперимент необходимо вследствие влияния случайного фактора u_1 поступающего на вход объекта. В связи с этим итоговые результаты, которые характеризуют значения относительных ошибок аппроксимации W_1 , W_2 , W_3 будут рассчитываться как среднее значение по результатам тридцати запусков работы алгоритма. Для полноты эксперимента объем выборки s , а также уровень помех k будут изменяться. Результаты данного вычислительного эксперимента приведены ниже и сведены в таблицу для удобства:

Таблица 1 – Результаты моделирования при различных параметрах s и k

Объем выборки s	Процент помехи k , %	Ошибка аппроксимации W_1 , %	Ошибка аппроксимации W_2 , %	Ошибка аппроксимации W_3 , %
100	0	41.69	12.87	10.13
	5	41.91	12.13	9.89
	10	43.25	13.85	12.41
300	0	23.19	6.95	3.27
	5	24.27	8.28	4.61
	10	24.72	8.57	6.98
500	0	17.81	6.01	2.17
	5	18.89	5.87	3.69
	10	19.56	7.62	6.44
700	0	15.21	3.41	1.67
	5	15.49	3.95	3.3
	10	17.25	7.9	6.25
900	0	14.02	4.12	1.37
	5	14.44	4.98	3.33
	10	15.06	7.54	4.12

Объем данных, представленный в таблице 1, удобнее визуализировать, для более очевидной интерпретации результатов. Построим график зависимости объема выборки s от значений ошибок аппроксимации (21), (22), (24). При этом

уровень помех $k = 5\%$. Полученная графическая зависимость представлена на рисунке 12 ниже:

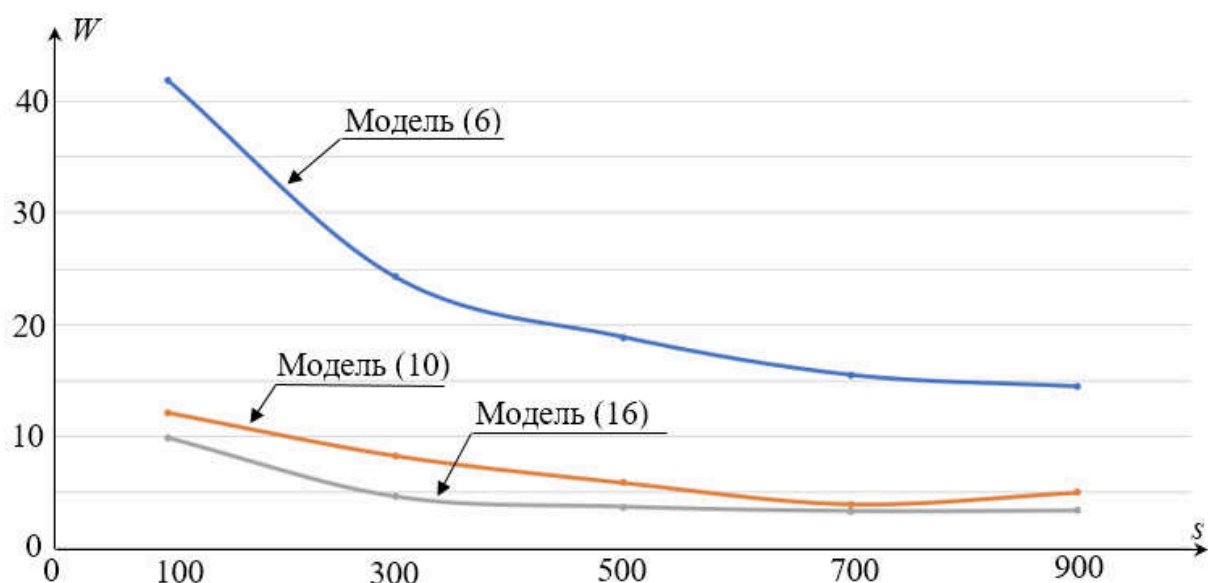


Рисунок 12 – Графическая интерпретация результатов из таблицы 1

3.5 Выводы по главе 3

В 3 главе были рассмотрены методики, позволяющие строить более точную непараметрическую модель исследуемого объекта благодаря «робастизации» непараметрической оценки или цензурированию исходных данных.

В ходе вычислительных экспериментов и интерпретации их результатов были выявлены следующие закономерности:

- при увеличении объема выборки точность аппроксимации повышается (рисунок 12);
- с ростом уровня помех качество моделирования падает (таблица 1);
- метод цензурирования данных дает более точные результаты относительно метода робастного оценивания функции регрессии, однако с увеличением объема выборки разница в результатах не столь велика (в худшем случае не больше 4%). В частности, данное утверждение может быть оправдано сравнительно небольшим количеством выбросов в данных, относительно объема

выборки. То есть, если в выборке, в среднем, среди тридцати экспериментов попадает 5 выбросов, которые в последствии удаляются, то это не является критичным для объема 100 точек и более;

- при использовании объекта одномерного или двумерного типа увеличения или уменьшения значений точности аппроксимации выявлено не было;

- при использовании робастной непараметрической оценки ремонт выборки выхода объекта лучше производить с использованием ее же оценочных значений. В таком случае, точность аппроксимации выше, если возникают ситуации, когда выбросы находятся по соседству друг с другом.

4 Задача оценки стоимости недвижимости

Одним из направлений выпускной квалификационной работы автора является непараметрическая идентификация дискретно-непрерывных процессов и в рамках данной тематики был рассмотрен вопрос оценки стоимости недвижимости. Задача решается с использованием выборки, содержащей в себе данные об однокомнатных квартирах города Красноярска. Проводятся вычислительные эксперименты, показывающие высокую точность полученной оценки.

Необходимо отметить, что недвижимость оказывает значительное влияние на жизнь человека и выступает в качестве ценного экономического ресурса. В связи с этим, активно ведутся разработки в области прогнозирования ее стоимости. Например, в работах [48, 49] рассматривается вопрос построения параметрической модели для оценки стоимости жилья. Впоследствии, к моделям применяются регрессионный и факторный анализ для того, чтобы выявить значимые или незначимые факторы, влияющие на конечное значение выхода модели. В данной работе решено было использовать иной подход к оцениванию стоимости жилья – непараметрический.

Стоимость квартиры зависит от большого числа факторов и имеет стохастический характер, что обуславливается экономической обстановкой в целом. В связи с этим достаточно сложно иметь представление о конечной стоимости жилья на данный момент. Когда перед человеком возникает задача покупки квартиры, то, в первую очередь, значительную роль играет бюджет, которым он располагает. Принимая во внимание проблему сложного прогнозирования цены, человек, который далек от понимания нюансов поведения рынка стоимости жилья, далеко не сразу сможет сориентироваться и подобрать оптимальный для него вариант. Именно для таких случаев прогнозирование стоимости квартиры и предоставление информации в доступном виде, с учетом личных потребностей и желаний клиента, является актуальной задачей.

Для решения различного рода задач, связанных с прогнозированием цены или предсказания переменных какого-либо процесса, явления, пользуются методами регрессионного анализа. Такие методы подразумевают изучение зависимости выходных переменных от входных. Важной проблемой в данном вопросе является выбор параметрического, либо непараметрического подхода к оцениванию регрессионной зависимости. Первый случай применим, если может быть задана структура объекта исследуемого процесса или явления, то есть, она представлена в виде известного функционального выражения, описываемого конечным набором параметров. Второй же случай используют при недостатке априорной информации, когда объект представлен в виде «черного ящика» и нет никакой возможности знать его структуру наверняка.

Помимо этого, непараметрический анализ позволяет осуществлять прогнозирование новых наблюдений. Чаще всего, мы уже обладаем каким-то набором данных, но, к сожалению, далеко не всегда они представляют полную картину процесса или явления. Так или иначе, вполне возможно спрогнозировать необходимое нам значение на основе имеющейся у нас априорной информации, которая, пусть даже, описывает процесс или явление не полностью. Результаты параметрических методов в данном вопросе являются слишком ограничительными для получения разумных объяснений наблюдаемых явлений, в то время как непараметрические методы дают лучшие результаты [28].

Также хочется добавить, что параметрические системы, связанные с прогнозированием, чаще всего являются сложно реализуемыми, особенно в условиях малой априорной информации. Их недостаток заключается в громоздком математическом описании, что в свою очередь усложняет процесс математического моделирования и проектирование адекватной модели исследуемого процесса или явления. Поэтому применить непараметрическую оценку, для которой нет необходимости знать точную структуру модели, будет более оптимальным решением.

4.1 Постановка задачи

В данной главе рассмотрены вопросы построения непараметрической модели и прогнозирования стоимости жилых объектов.

Вычислительный эксперимент, который будет описан далее, опирается на данные, содержащие информацию о стоимости однокомнатных квартир в городе Красноярске, которые включают в себя 1358 объектов. Выборка не содержит пропусков и выбросов. Данные представлены следующим признаковым описанием квартир: общая площадь – u_1 , площадь кухни – u_2 , район – u_3 , этаж – u_4 , материал стен – u_5 и планировка – u_6 . Выходной переменной x является цена квартиры. Следует упомянуть, что в роли количественных переменных здесь выступают площадь кухни, жилая площадь и цена, а остальные признаки являются номинальными.

Выборку, описанную выше, можно охарактеризовать как процесс дискретно-непрерывного типа, описание которого приведено в работе ранее (рисунок 2 параграфа 2.2).

Для построения регрессионной модели была выбрана непараметрическая оценка Надарая-Ватсона, вид которой в одномерном случае представлен ниже:

$$\hat{x} = \frac{\sum_{i=1}^s x_i \Phi\left(\frac{u - u_i}{c_s}\right)}{\sum_{i=1}^s \Phi\left(\frac{u - u_i}{c_s}\right)}, \quad (25)$$

где $\{x_i, u_i, i=1, \dots, s\}$ – исходная выборка наблюдений;

c_s – параметр размытости ядра;

$\Phi(\bullet)$ – ядерная функция;

s – объем выборки.

4.2 Непараметрическая модель

Ядро $\Phi(\bullet)$ для непараметрической оценки (25) решено было взять параболическим, так как оно дает более точные результаты относительно стандартных треугольного и прямоугольного. Вид параболического ядра представлен ниже:

$$\Phi^{(\kappa)}(z) = \begin{cases} (0.75 \cdot (1 - |z|)^2, & |z| < 1, \\ 0, & |z| \geq 1, \end{cases} \quad (26)$$

где $z = \frac{u - u_i}{c_s}$.

Ядро вида (26) используется только для входных количественных признаков. Для номинальных признаков ядро принимает следующий вид:

$$\Phi^{(H)}(z) = \begin{cases} 1, & u = u_i, \\ 0, & u \neq u_i. \end{cases} \quad (27)$$

На основании приведенных выше формул модель, используемая в данной работе, принимает вид:

$$\hat{x} = \frac{\sum_{i=1}^s x_i \prod_{j=1}^2 \Phi^{(\kappa)}\left(\frac{u_j - u_{ji}}{c_s}\right) \cdot \prod_{j=3}^6 \Phi^{(H)}(u_j - u_{ji})}{\sum_{i=1}^s \prod_{j=1}^2 \Phi^{(\kappa)}\left(\frac{u_j - u_{ji}}{c_s}\right) \cdot \prod_{j=3}^6 \Phi^{(H)}(u_j - u_{ji})}. \quad (28)$$

Точность полученной модели оценивалась с помощью модульно критерия:

$$W = \frac{1}{s} \sum_{i=1}^s |x_i - \hat{x}_i|, \quad (29)$$

где \hat{x}_i – итоговые значения выхода модели (28);

x_i – объект (в данном случае представленный выборкой наблюдений однокомнатных квартир города Красноярска).

4.3 Вычислительный эксперимент

На первом этапе необходимо найти оптимальный параметр размытости c_s для непараметрической оценки с учетом всех факторов, влияющих на стоимость жилья. Настраивать его будем в режиме скользящего экзамена.

Ниже, на рисунке 13, представлен график зависимости ошибки, рассчитанной по формуле (29) от значения параметра размытости c_s .

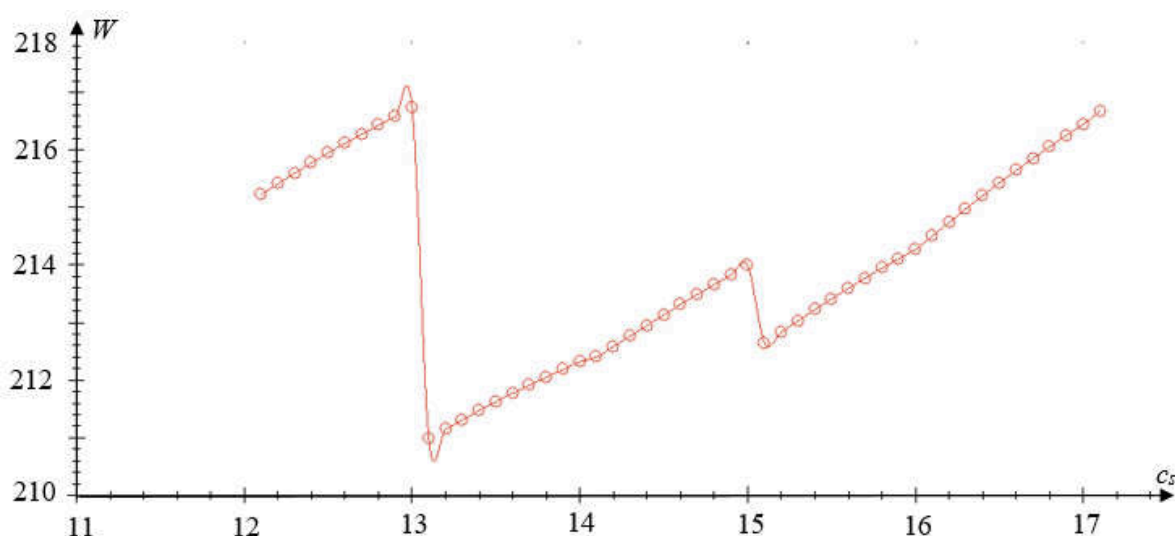


Рисунок 13 – Зависимость значения ошибки аппроксимации (29) от величины параметра размытости

Можно заметить, что минимум ошибки достигается в интервале примерно от 13 до 13.5. Более наглядная интерпретация результатов, представленных на рисунке 13 приведена в таблице ниже:

Таблица 2 – Численные значения ошибок моделирования при каждом значении параметра размытости

Значение c_s	Значение ошибки моделирования W , тыс. руб.
13	216.73
13.1	210.98
13.2	211.15
13.3	211.31
13.4	211.47
13.5	211.62

По результатам, приведенным в таблице 2, можно с уверенностью сказать, что оптимальное значение $c_s = 13.1$, так как для него ошибка моделирования минимальна.

На основании приведенных выше результатов, построим непараметрическую оценку (28) для оптимального значения параметра размытости $c_s = 13.1$, учитывая все факторы, влияющие на стоимость квартиры. Результаты моделирования представлены ниже, на рисунке 14.

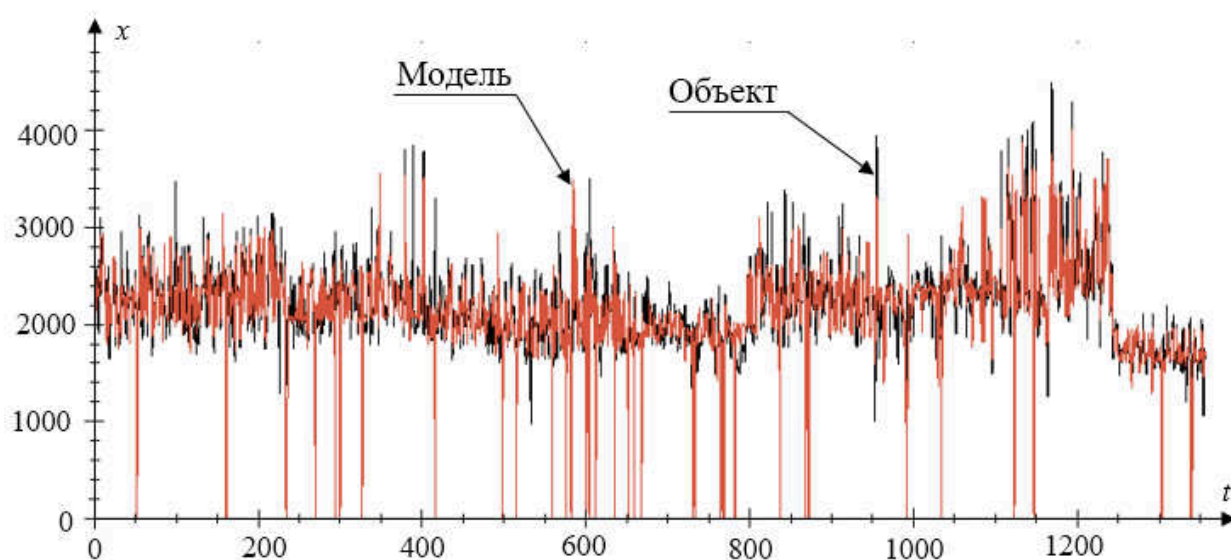


Рисунок 14 – Зависимость выхода модели и объекта от такта времени

По результатам, представленным на рисунке 14 видно, что в некоторых случаях непараметрическая оценка не может быть рассчитана (нулевые значения графика модели). В частности, значение стоимости неизвестно для 37 жилых объектов. Это объясняется тем, что в исходной выборке недостаточно информации для прогнозирования, а также тем, что фактор u_1 накладывает ограничение на точки, входящие под колокол ядерной функции.

Проблема, описанная выше, послужила поводом для дальнейшего изучения оценки (28). Найдем оптимальный параметр размытости c_s аналогично способу, описанному в ранее, но без учета фактора u_1 , который, как оказалось, сильно влияет на конечные результаты выхода модели. Итоговые результаты представлены ниже на рисунке 15.

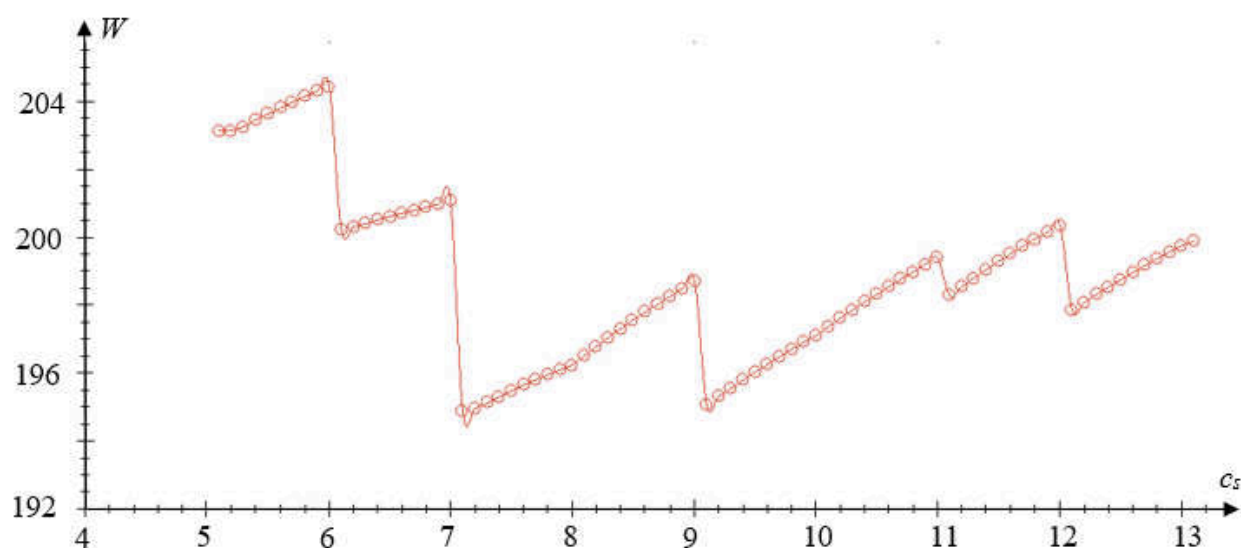


Рисунок 15 – Зависимость значения ошибки аппроксимации (29) от величины параметра размытости

По графику, представленному на рисунке 15 видно, что оптимальное значение параметра размытости равно $c_s = 7.1$ при значении модульного критерия (29) равного $W=194.87$.

Основываясь на полученных результатах, построим непараметрическую оценку при оптимальном значении параметра размытости. Ниже, на рисунке 16 продемонстрированы итоговые результаты моделирования.

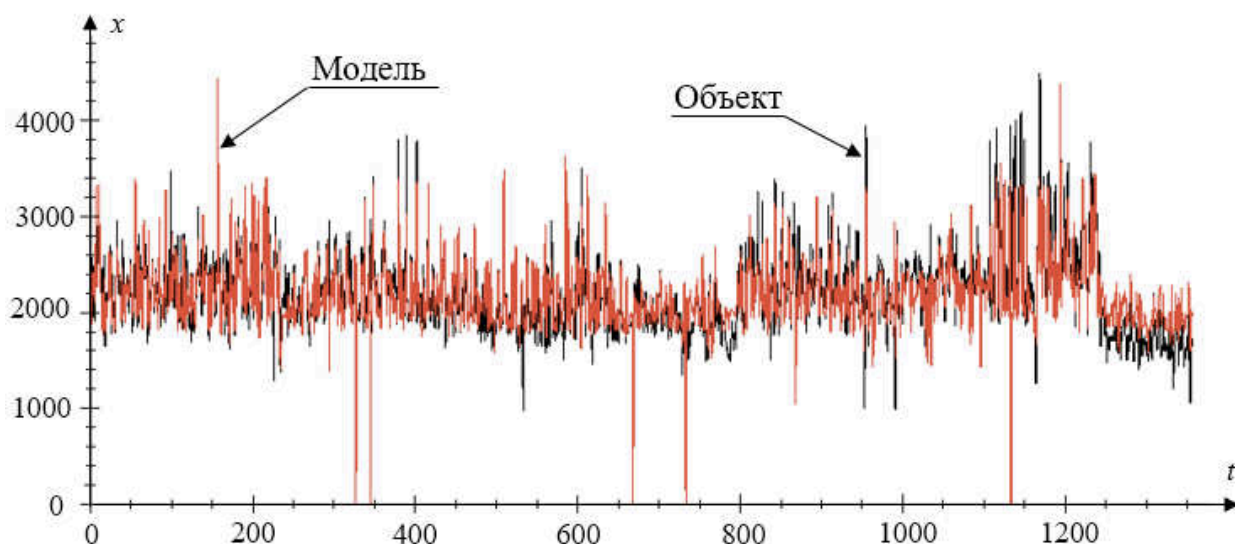


Рисунок 16 – Зависимость выхода модели и объекта от тактов времени

Таким образом, исходя из приведенных выше результатов видно, что непараметрическая модель, построенная без учета фактора u_1 , точнее описывает объект. В данном случае неизвестно значение стоимости только лишь для 6 жилых объектов.

4.4 Информационная система

Представленная оценка (28), где учитывается фактор u_1 , легли в основу создания информационной системы [50], в которой предусмотрена возможность задавать параметры квартиры по своему усмотрению и получить в результате стоимость жилья относительно своих предпочтений на основании обучающей выборки.

Для демонстрации примера работы информационной системы возьмем 2 район, общую площадь квартиры равную 42 м^2 , площадь кухни 8 м^2 , этаж 1, материал стен 2, 4 планировку. В результате рассчитана итоговая стоимость равная 2 миллионам 269 тысячам рублей. Результат вычислений в интерфейсе разработанной информационной системы представлен на рисунке 17:

Оценка стоимости жилья г. Красноярск

Параметры желаемой квартиры:

Общая площадь	от 42 до 42	Материал стен	от 2 до 2
Площадь кухни	от 8 до 8	Планировка	от 4 до 4
Район	от 2 до 2		
Этаж	от 1 до 1	Рассчитанная стоимость	2269

Рассчитать

Рисунок 17 – Интерфейс информационной системы

К сожалению, возникают ситуации, при которых значение цены может быть не определено. Как и в предыдущем параграфе они обоснованы недостаточным объемом выборки данных. Описанную проблему возможно решить, используя большее количество квартир в выборке наблюдений.

Клиент, выбирающий квартиру, не всегда имеет возможность знать параметры своего будущего жилья наверняка. Возникают ситуации, при которых параметры квартиры должны быть варьируемыми, то есть должны быть заданы интервально, в зависимости от желаний клиента, чтобы не ограничиваться, к примеру, только в 9 квадратах кухни, а иметь возможно узнать примерную стоимость квартир с кухнями от 9 м² до 13м². И такая возможность учитывается в системе.

Допустим, из предыдущего примера, нас устраивает 2 район, 2 материал стен и 4 планировка, однако хотелось бы узнать стоимость жилья с площадью от 42 м² до 56 м², площадью кухни от 8м² до 15м² и рассмотреть квартиры не только на 1 этаже, а также и на втором. Таким образом, учитывая все наши желания, итоговая стоимость получилась ниже и стала равна 2 миллионам 150 тысячам рублей.

4.5 Вывода по главе 4

Таким образом, с использованием непараметрического алгоритма была разработана информационная система для прогнозирования стоимости жилых объектов, где предусмотрена возможность настройки параметров квартиры точно и интервально.

Оказалось, что на используемую непараметрическую оценку оказывает влияние фактор района, в котором находится квартира. Без учета данного фактора была построена более точная оценка стоимости жилых объектов, что также было подтверждено значениями ошибок аппроксимации.

Результаты и выводы, описанные в данной главе, были представлены на международной научной конференции, а также опубликованы в сборнике трудов конференции [51].

Помимо всего прочего, автор хочет выразить особую благодарность Иконникову Олегу Александровичу за предоставление данных о характеристиках и стоимости однокомнатных квартир города Красноярска.

ЗАКЛЮЧЕНИЕ

Цель данной работы заключалась в повышении точности идентификации дискретно-непрерывных процессов одномерного и двумерного типа с выборкой наблюдений, содержащей выбросы.

Результаты вычислительных экспериментов на практике подтвердили целесообразность использования описанных в работе алгоритмов. Точность аппроксимации увеличилась более чем в два раза как при использовании робастного алгоритма идентификации, так и при цензурировании выборки наблюдений. При том количестве выбросов, что существовало в рамках вычислительного эксперимента (в среднем не больше шести) значительной разницы в точности построения модели без выбросов выявлено не было. Помимо этого, были рассмотрены объекты двух типов, одномерный и двумерный. При увеличении числа входных переменных объекта, ухудшения в точности аппроксимации не обнаружено.

После исключения выбросов из выборки наблюдений был проведен ремонт данных. Для модели, построенной с использованием робастного аналога непараметрической оценки, лучшим решением, с точки зрения повышения точности, стало использование значений робастной модели в качестве восстановления точек, являющихся выбросами. При использовании не робастной модели были выявлены ситуации, когда точность аппроксимации объекта с восстановленными данными меньше, а именно наличие в выборке двух или более выбросов, находящихся по соседству друг с другом.

Далее, основываясь на результатах, описанных выше, была построена непараметрическая модель с использованием реальной выборки наблюдений. В ходе моделирования, было выявлено ухудшение точности, в связи с ограничениями, которые накладывает фактор входной переменной u_1 на точки, попадающие под колокол ядерной функции. После выявленной проблемы была построена новая непараметрическая оценка без учета данного фактора, что несколько увеличило точность моделирования объекта исследования.

Разобранные алгоритмы идентификации систем служат основой при моделировании процессов. На базе проведенных вычислительных экспериментов, в дальнейшем, можно конструировать модели более сложных по своей структуре и составу объектов, процессов или явлений.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ваганов, М. А. Многоканальный спектральный прибор для диагностики жидкостного ракетного двигателя / М. А. Ваганов, О. Д. Москалец, С. В. Кулаков // Информационно-управляющие системы. – 2013. – № 1 (62). – С. 2–6.
2. Мухин, С. В. Перспективы развития информационно-измерительных и управляющих систем для испытания жидкостного ракетного двигателя на стенде химзавода – филиала ОАО «КРАСМАШ» / С. В. Мухин, А. В. Ребенков // Решетневские чтения: материалы междунар. науч. конф. / Изд-во Сиб. гос. аэрокосмич. ун-та, 2010. – С. 261–266.
3. Abbott, B. Observation of Gravitational Waves from a Binary Black Hole Merger / B. Abbott [et al.]. // Physical Review Letters. – 2016. – № 116 (6). – P. 061102.
4. Советов, Б. Я. Моделирование систем : Учебник для вузов / Б. Я. Советов, С. А. Яковлев. – 3-е изд., перераб. и доп. – Москва : Высш. шк., 2001. – С. 6.
5. Харин, Ю. С. Основы имитационного и статистического моделирования : Учебное пособие / Ю. С. Харин [и др.]. – Минск : Дизайн ПРО, 1997. – С. 5.
6. Шеннон, Р. Имитационное моделирование – искусство и наука. / Р. Шеннон. – Москва : Мир, 1978. – 418 с.
7. Самарский, А. А. Математическое моделирование: Идеи. Методы. Примеры. / А. А. Самарский, А. П. Михайлов. – 2-е изд., испр. – Москва : ФИЗМАТЛИТ, 2005. – С. 7.
8. Медведев, А. В. Основы теории адаптивных систем : монография / А. В. Медведев. – Красноярск : Сибирский гос. аэрокосмический ун-т им. акад. М. Ф. Решетнева, 2015. – 524 с.
9. Цыпкин, Я. З. Основы информационной теории идентификации. / Я. З. Цыпкин. – Москва : Наука, 1984. – 320 с.
10. Murphy, K. P. Machine Learning: A Probabilistic Perspective. / K. P. Murphy. – Cambridge : MIT Press, 2012. – P. 2.

11. Вапник, В. Н. Восстановление зависимостей по эмперическим данным. / В. Н. Вапник. – Москва : Наука, 1979. – 448 с.
12. Flach, P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. / P. Flach. – Cambridge : Cambridge University Press, 2012. – 367 p.
13. Billings, J. M. Machine Learning Applications to Resting-State Functional MR Imaging Analysis. / J. M. Billings [et al.]. // Neuroimaging Clinics of North America. – 2017. – P. 609–620.
14. Mahdavejad, M. S. Machine learning for Internet of Things data analysis: A survey. / M. S. Mahdavejad [et al.]. // Digital Communications and Networks. – 2017. – P. 1-56.
15. Бокс, Г. Анализ временных рядов: Прогноз и управление. / Г. Бокс, Г. Дженкинс. – Москва : Мир, 1974. 406 с.
16. Fahrmeir, L. Regression: Models, Methods and Applications / L. Fahrmeir [et al.]. Munich – 2013. – 698 p.
17. Ивахненко, А. Г. Долгосрочное прогнозирование и управление сложными системами. / А. Г. Ивахненко. – Киев : Техника, 1975. – С. 6.
18. Абрамов, А. В. Прогнозирование валютного курса EUR/USD с использованием нейронных сетей / А. В. Абрамов // Ученые записки. Электронный научный журнал Курского государственного университета. – 2012. – №4. – С. 71–76.
19. Liang, C. Developing accident prediction model for railway level crossings. / C. Liang [et al.]. // Safety Science. – № 101 – 2018, P. 48-59.
20. Эйкхофф, П. Основы идентификации систем управления. / П. Эйкхофф. – Москва : Мир, 1975. – 686 с.
21. Гроп, Д. Методы идентификации систем. / Д. Гроп. – Москва : Мир, 1979. – С. 11.
22. Денисов, М. А. About parametric identification algorithms of discrete-continuous processes / М. А. Денисов, Е. А. Чжан // Сибирский журнал науки и технологий. – 2018. – № 4. (8) – С. 727-735.

23. Корнеева, А. А. О параметрическом моделировании стохастических объектов / А. А. Корнеева, Е. А. Чжан // Вестник СибГАУ им. М. Ф. Решетнева. – 2013. – № 2. – С. 39–42.
24. Beltran-Carbajal, F. On-line parametric estimation of damped multiple frequency oscillations. / F. Beltran-Carbajal, G. Silva-Navarro, L. G. Trujillo-Franco // Electric Power Systems Research – 2018. – № 154. – P. 423–432.
25. Герасимов, А. Н. Параметрические и непараметрические методы в медицинской статистике / А. Н. Герасимов, Н. И. Морозова // Эпидемиология и вакцинопрофилактика. – 2015. – Т. 14. – № 5. – С. 6–12.
26. Memon, A. G. Parametric based economic analysis of a trigeneration system proposed for residential buildings. / A. G. Memon, R. A. Memon // Sustainable Cities and Society. – 2017. – № 34. – P. 144–158.
27. Катковник, В. Я. Непараметрическая идентификация и сглаживание данных: метод локальной аппроксимации. / В. Я. Катковник. – Москва : Главная редакция физико-математической литературы, 1985. – С. 6.
28. Härdle, W. Applied nonparametric regression. / W. Härdle. – Cambridge : Cambridge university press, 1989. – 434 p.
29. Агафонов, Е. Д. Непараметрическая модель в задаче прогнозирования мощности ветряных электрических установок / Е. Д. Агафонов, Е. С. Мангалова, О. В. Шестернева // Вестник СибГАУ им. М. Ф. Решетнева. – 2013. – № 2. – С. 4–9.
30. Соколов, М. И. Непараметрические модели оценивания показателей эффективности агрегатов системы терморегулирования космических аппаратов / М. И. Соколов // Вестник СГАУ. – 2007. – № 1. – С. 81–89.
31. Большаков, А. А. Методы обработки многомерных данных и временных рядов : Учебное пособие для вузов / А. А. Большаков, Р. Н. Каримов. – Москва : Горячая линия-Телеком, 2007. – 522 с.
32. Гайдышев, И. Анализ и обработка данных: специальный справочник. / И. Гайдышев. – Санкт-Петербург : Питер, 2001. – 752 с.

33. Джексон, П. Введение в экспертные системы. / П. Джексон. – Москва : Вильямс, 2001. – 397 с.
34. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний. / Н. Г. Загоруйко. – Новосибирск : Изд-во Ин-та математики, 1999. – С 7-8.
35. Корнеева, А. А. Об анализе данных в задаче идентификации статических систем / А. А. Корнеева, Н. А. Сергеева // Вестник СибГАУ им. М. Ф. Решетнева. – 2012. – №. 5. – С. 49–54.
36. Huber, P. J. Robust statistics. / P. J. Huber // Statistics. – 2004. – № 60. – P. 1-11.
37. Maronna, R. Robust Statistics: Theory and Methods. / R. Maronna, R. Martin, V. Yohai. – Chichester : John Wiley & Sons, 2006. – 978 p.
38. Шуленин, В. П. Робастные методы математической статистики. / В. П. Шуленин. – Томск: Изд-во НТЛ, 2016. – 260 с.
39. Харин, Ю. Робастная статистика и ее применение / Ю. Харин // Наука и инновации. – 2010. – Т. 8. – № 90. – С. 22-23
40. Горяинов, В. Б. Робастное оценивание в авторегрессионной модели со случайным коэффициентом / В. Б. Горяинов, С. Ю. Ермаков // Наука и образование: научное издание МГТУ им. Н.Э. Баумана. – 2016. – №9. – С. 111-122
41. Тихонов, В. И. Выбросы случайных процессов. / В. И. Тихонов. – Москва : Наука, 1970. – 392 с.
42. Cohen, W. W. Fast effective rule induction / W. W. Cohen // Proceedings of the twelfth international conference on machine learning. – 1995. – P. 115-123.
43. Brodley, C. E. Identifying mislabeled training data / C. E. Brodley, M. A. Friedl // Journal of artificial intelligence research. – 1999. – № 11. – P. 131-167.
44. Рубан, А. И. Методы анализа данных : учебник / А. И. Рубан. – Красноярск: Сиб. Федер. ун-т, 2012. – С. 153.
45. Надарая, Э. А. Непараметрические оценки плотности вероятности и кривой регрессии. / Э. А. Надарая. – Тбилиси : Изд-во Тбил. ун-та, 1983. – 194 с.

46. Кирик, Е. С. Моделирование и оптимизация робастных оценок функций по наблюдениям / Е. С. Кирик // Вычислительные технологии. – 2001. – Т. 6. – С. 351-355.

47. Корнеева, А. А. Непараметрические модели и алгоритмы управления для многомерных систем с запаздыванием : дис. канд. тех. наук. : 05.13.01. / Корнеева Анна Анатольевна. – Красноярск, 2014. – 176 с.

48. Реннер, А. Г. Моделирование стоимости жилья на вторичном рынке жилья / А. Г. Реннер, О. И. Стебунова // Вестник Оренбургского государственного университета. – 2005. – № 10-1. – 179-182.

49. Сенашов, С. И. Информационная система оценки стоимости квартир на вторичном рынке жилья как инструмент управления инвестициями / И. С. Сенашов, Н. Ю. Юферова, Е. В. Сурнина // Вестник СибГАУ им. М. Ф. Решетнева. – 2009. – № 4. – С. 219-222.

50. Свид. 2017663876 Российская Федерация. Свидетельство об официальной регистрации программы для ЭВМ. Информационная система оценки стоимости недвижимости. / М. А. Денисов, А. А. Корнеева; заявитель и правообладатель ФГАОУ ВО СФУ (RU). – №2017660352; заявл. 16.10.17; опубл. 13.12.17, Реестр программ для ЭВМ. – 1 с.


Федеральное государственное автономное образовательное учреждение
высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
Базовая кафедра интеллектуальных систем управления

УТВЕРЖДАЮ

Заведующий кафедрой

 Ю.Ю. Якунин

«11» июня 2018 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 Системный анализ и управление

Идентификация многомерных дискретно-непрерывных процессов по выборке
наблюдений с выбросами

Руководитель



подпись, дата

Выпускник



подпись, дата

доцент, к.т.н.

должность, ученая степень

А.А. Корнеева

инициалы, фамилия

М.А. Денисов

инициалы, фамилия

Красноярск 2018